华为职业认证通过者权益

通过**任一项**华为职业认证，您即可在华为在线学习网站(*http://learning.huawei.com/cn*）享有如下特权：

- **1、华为E-learning 课程学习**
  - **内容**：**所有华为职业认证E-Learning课程**，扩展您在其他技术领域的技术知识
  - **方式**：请提交您的"华为账号"和注册账号的"email地址"到 *Learning@huawei.com 申请权限。*
- **2、华为培训教材下载**
  - **内容：华为职业认证培训教材+华为产品技术培训教材，**覆盖企业网络、存储、安全等诸多领域
  - **方式**：登录*华为在线学习网站*，进入"*华为培训->面授培训*"，在具体课程页面即可下载教材。
- **3、华为在线公开课(LVC)优先参与**
  - **内容**：企业网络、UC&C、安全、存储等诸多领域的职业认证课程，华为讲师授课，开班人数有限
  - **方式**：开班计划及参与方式请详见LVC排期：
    http://support.huawei.com/learning/NavigationAction!createNavi#navi[id]=_16
- **4、学习工具 eNSP**
  - *eNSP (Enterprise Network Simulation Platform)*, 是由华为提供的免费的、可扩展的、图形化网络仿真工具。主要对企业网路由器和交换机进行硬件模拟，完美呈现真实设备实景；同时也支持大型网络模拟，让大家在没有真实设备的情况下也能够进行实验测试。
- 另外，华为建立了知识分享平台 华为认证论坛。您可以在线与华为技术专家交流技术，与其他考生分享考试经验，一起学习华为产品技术。（http://support.huawei.com/ecommunity/bbs/list_2247.html ）



华为认证系列教程

HCNP-Storage CUSN

构建统一存储网络



华为技术有限公司

华为认证系列教程

HCNP-Storage CUSN 构建统一存储网络

第2.0版本

华为认证系统介绍

依托华为公司雄厚的技术实力和专业的培训体系，华为认证考虑到不同客户对 ICT 技术不同层次的需求，致力于为客户提供实战性、专业化的技术认证。 根据 ICT 技术的特点和客户不同层次的需求，华为认证为客户提供面向十三个方向的四级认证体系。

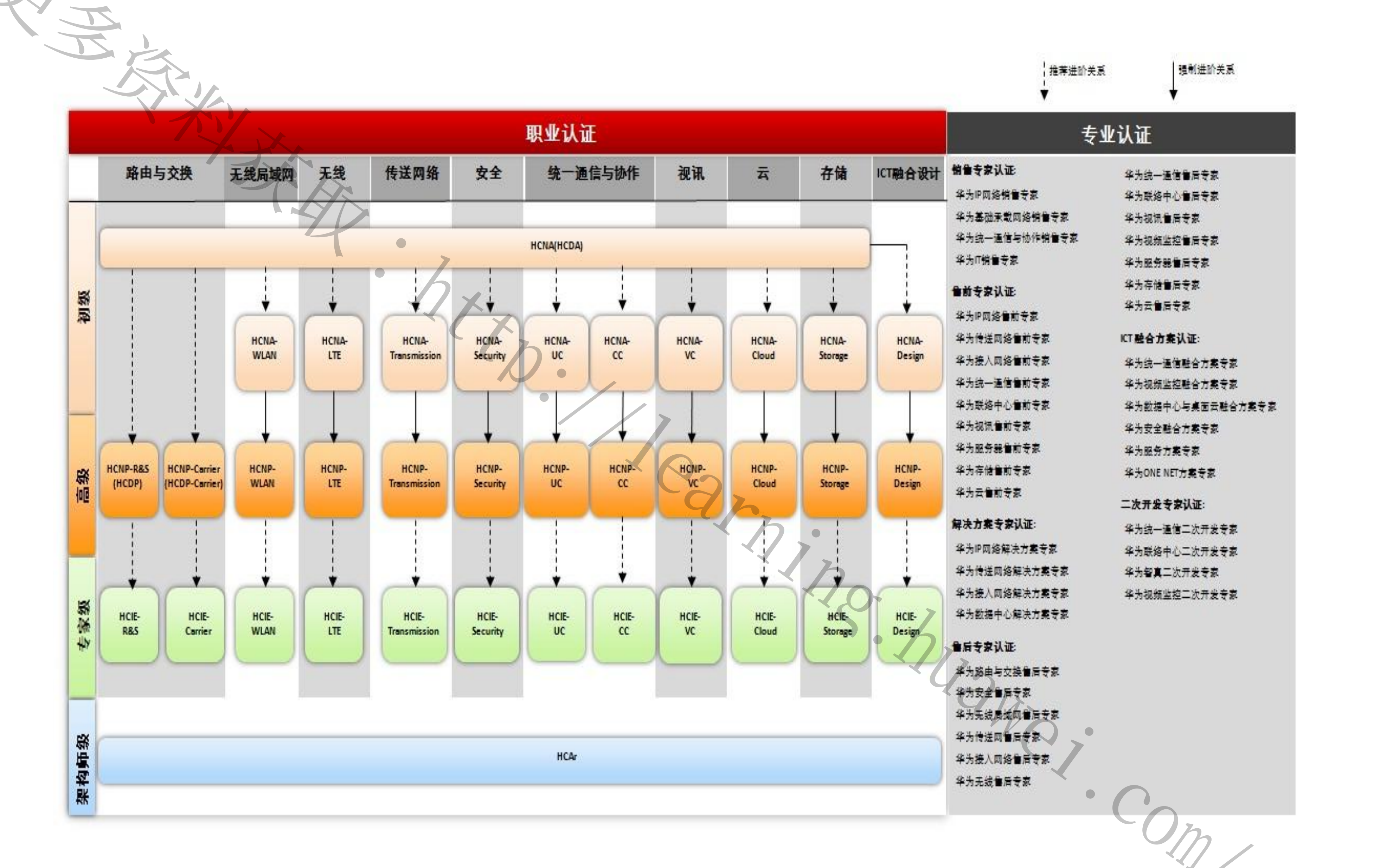
HCNA-Storage（Huawei Certified Network Associate-Storage，华为网络存储工程师认证）主要面向网络存储维护工程师，以及准备参加 HCNA-Storage 认证考试的人员；希望掌握 SAN 存储系统与网络基本原理和华为 SAN 存储阵系统管理、部署与维护能力的人员。HCNA 认证在内容上涵盖存储基础知识、RAID 技术与应用、存储网络技术与应用、华为存储产品与解决方案、存储系统管理和基本配置、存储主机连接与多路径配置、SAN 网络与存储系统日常维护。

HCNP-Storage (Huawei Certified Network Professional-Storage，华为认证网络存储资深工程师)主要面向企业级网络存储维护工程师、专家工程师以及希望系统深入掌握 SAN 存储、统一存储、数据保护技术与部署的人员。

HCNP-Storage 包括 CUSN（Constructing Unifying Storage Network，构建统一存储网络）、CBDS（Constructing Big Data Solution 构建大数据解决方案）、CDPS（Constructing Data Protection System 构建数据保护系统）三个部分。内容上涵盖 SAN、NAS 统一存储原理、架构和组件，存储数据处理与通信协议（SCSI、FC、iSCSI）原理及应用，存储系统数据可靠性与业务连续性保障技术存储与主流 OS 平台连接与应用，存储网络冗余技术及应用，SAN、集群 NAS 网络规划与方案部署，虚拟快照、LUN Copy、复制的原理和部署，网络存储虚拟化技术及应用，存储虚拟化系统原理、部署和异构资源管理，备份网络及备份恢复技术及应用，华为数据保护方案构建、部署与管理，华为数据容灾方案及典型应用场景，华为存储系统、网络、方案的故障诊断与处理方法。

HCIE-Storage（Huawei Certified Internetwork Expert--Storage，华为认证互联网络存储专家）旨在培养能够熟练掌握各种存储技术；精通 IT 存储方案设计、部署和运维管理以及诊断和故障排除。

华为认证协助您打开行业之窗，开启改变之门，屹立在 ICT 世界的潮头浪尖！



推荐进阶关系
强制进阶关系
职业认证
专业认证
路由与交换
无线局域网
无线
传送网络
安全
统一通信与协作
视讯
云
存储
ICT融合设计
初级
高级
专家级
架构师级
HCNA(HCDA)
HCNA-WLAN
HCNA-LTE
HCNA-Transmission
HCNA-Security
HCNA-UC
HCNA-CC
HCNA-VC
HCNA-Cloud
HCNA-Storage
HCNA-Design
HCNP-R&S (HCDP)
HCNP-Carrier (HCDP-Carrier)
HCNP-WLAN
HCNP-LTE
HCNP-Transmission
HCNP-Security
HCNP-UC
HCNP-CC
HCNP-VC
HCNP-Cloud
HCNP-Storage
HCNP-Design
HCIE-R&S
HCIE-Carrier
HCIE-WLAN
HCIE-LTE
HCIE-Transmission
HCIE-Security
HCIE-UC
HCIE-CC
HCIE-VC
HCIE-Cloud
HCIE-Storage
HCIE-Design
HCAr
销售专家认证:
华为IP网络销售专家
华为基础承载网络销售专家
华为统一通信与协作销售专家
华为IT销售专家
售前专家认证:
华为IP网络售前专家
华为传送网络售前专家
华为接入网络售前专家
华为统一通信售前专家
华为数据中心售前专家
华为视讯售前专家
华为服务器售前专家
华为存储售前专家
华为云售前专家
解决方案专家认证:
华为IP网络解决方案专家
华为传送网络解决方案专家
华为接入网络解决方案专家
华为数据中心解决方案专家
售后专家认证:
华为路由与交换售后专家
华为安全售后专家
华为无线局域网售后专家
华为传送网售后专家
华为接入网络售后专家
华为无线售后专家
华为统一通信售后专家
华为数据中心售后专家
华为视讯售后专家
华为视频监控售后专家
华为服务器售后专家
华为存储售后专家
华为云售后专家
ICT融合方案认证:
华为统一通信融合方案专家
华为视频监控融合方案专家
华为数据中心与桌面云融合方案专家
华为安全融合方案专家
华为服务方案专家
华为ONE NET方案专家
二次开发专家认证:
华为统一通信二次开发专家
华为数据中心二次开发专家
华为智真二次开发专家
华为视频监控二次开发专家

前言

简介

HCNP-Storage 认证定位于 IT 领域信息存储高级工程师或存储方案专家能力构建。

本书为 HCNP-Storage 认证培训教程，适用于华为认证网络存储资深工程师以及准备参加 HCNP-Storage 认证考试的人员，通过 HCNP-Storage 认证，将证明您对信息存储系统、网络有全面深入的理解，掌握存储系统和网络（SAN、NAS、灾备系统）的通用技术，并具备独立完成信息存储、数据保护等综合解决方案的部署、运维和管理。

内容描述

本书是 HCNP-Storage-CUSN（Huawei Certified Network Professional - Constructing Unifying Storage Network 华为认证网络存储资深工程师－构建统一存储网络），用于指导学员学习 HCNP-Storage-CUSN 认证考试（H13-621-CHS）相关内容。共包含 8 个 Module，内容覆盖：SAN 存储架构、组件和存储数据处理与通信协议（SCSI、FC、iSCSI）原理及应用，存储数据可靠性与业务连续性保障技术，存储性能因素与调优，存储系统管理特性与应配置，存储与主流 OS 平台连接与应用，存储网络冗余技术及应用，虚拟快照、LUN Copy、复制的原理、规划与部署，SAN 存储系统与网络典型场景故障诊断思路和故障处理方法。

Module 1　统一存储协议：SAN 存储系统及结构、SCSI 协议及存储架构模型、FC 协议、SAS 协议、iSCSI 协议、NAS 系统及结构、CIFS 协议、NFS 协议。

Module 2　统一存储技术及应用：统一存储技术、RAID2.0 技术、Smart tier、Hyper Copy、Smart thin、Hyper Clone、Ultra APM、存储业务连接性方案。

Module 3　虚拟化存储网关系统部署与管理：存储虚拟化简介，存储虚拟化技术实现分类，基于各层的存储虚拟化，拟化存储网关产品架构与软硬件介绍，虚拟化存储网关功能特性介绍，快照技术及应用，复制技术及应用。

Module 4　统一存储与主机连接：存储与 UNIX 主机连接，多路径部署与管理，多路径故障处理。

Module 5　统一存储系统故障诊断：故障诊断原则，流程和方法，SAN 存储故障处理思路和方法，VIS 存储故障处理思路和方法。

Module 6　统一存储系统性能与优化：性能调优概述，性能指标，影响性能的关键因素及技术，性能诊断和调优，性能测试工具和方法，SAN 存储系统常见性能故障排除。

Module 7　统一存储方案规划与设计：规划原则及流程、主机层规划、网络层规划、存储层规划、规划案例。

Module 8　OceanStor 18000 存储系统安装及维护：硬件结构介绍，硬件安装和连线，勘测和包装拆卸，维护工具，兼容性基础。

最终掌握存储系统和网络（SAN）的通用技术及应用，并具备完成信息存储系统及 SAN 存储解决方案规划、部署、运维、管理能力。

读者必备知识背景

本课程为华为认证网络存储高级课程，阅读本书的读者应首先具备以下基本条件：

1、 了解存储技术和 SAN 和 NAS 架构与组成

2、 熟悉 SAN 存储系统配置与管理

3、 熟悉主流 OS 原理和安装部署，熟悉设备管理操作

4、 熟悉文件系统基本知识

5、 熟悉备份、容灾等数据保护概念

6、 有初步的备份项目实施经验

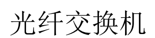
本书常用图标











光纤交换机　　以太网交换机　　存储系统　　主机

目 录

HC120920001 统一存储系统概述




HC120920001
统一存储系统概述
www.huawei.com
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



目标

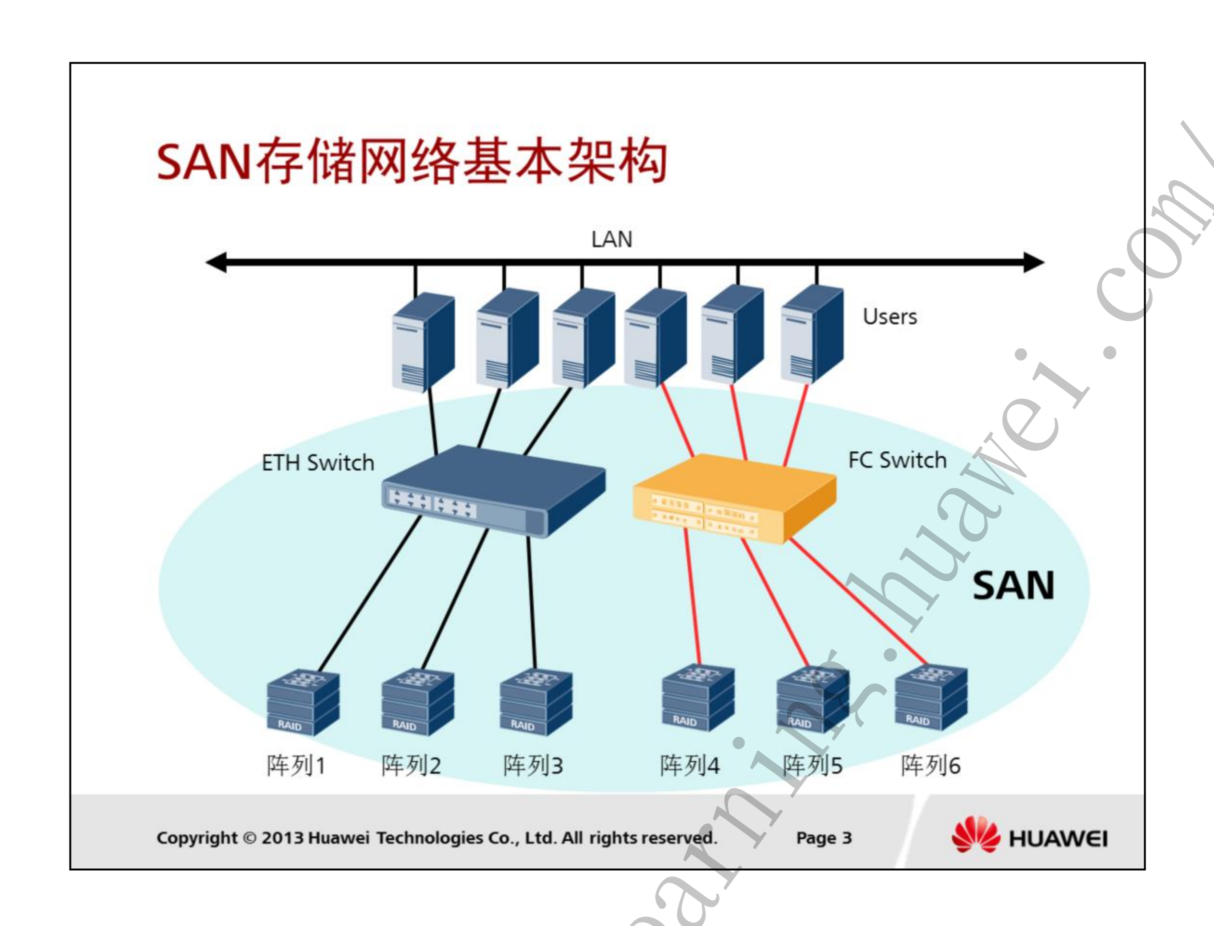
- 学完本课程后，您将能够:
  - 熟悉SAN存储系统的基本构架
  - 熟悉SAN存储系统的协议
  - 熟悉NAS存储系统的基本构架
  - 熟悉NAS存储系统的协议



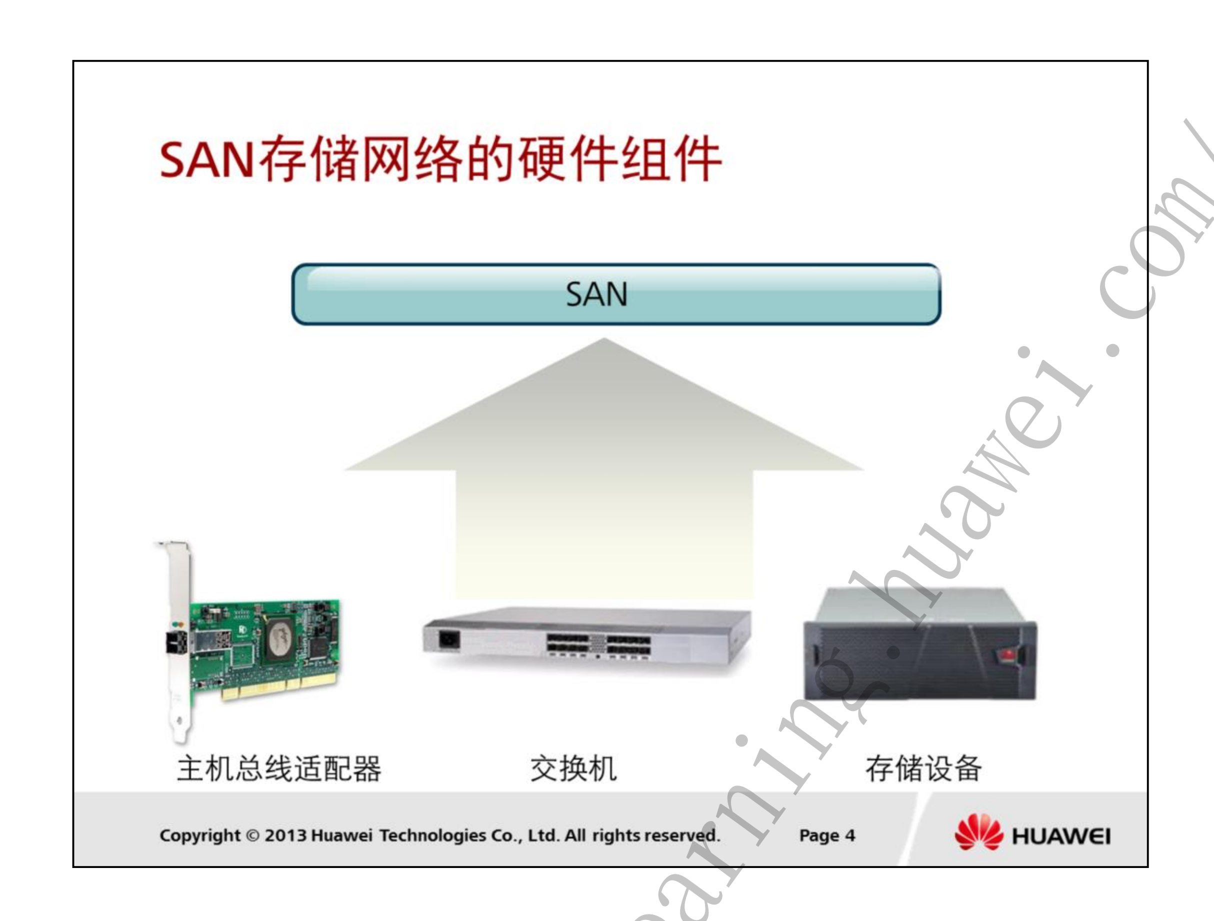


目录



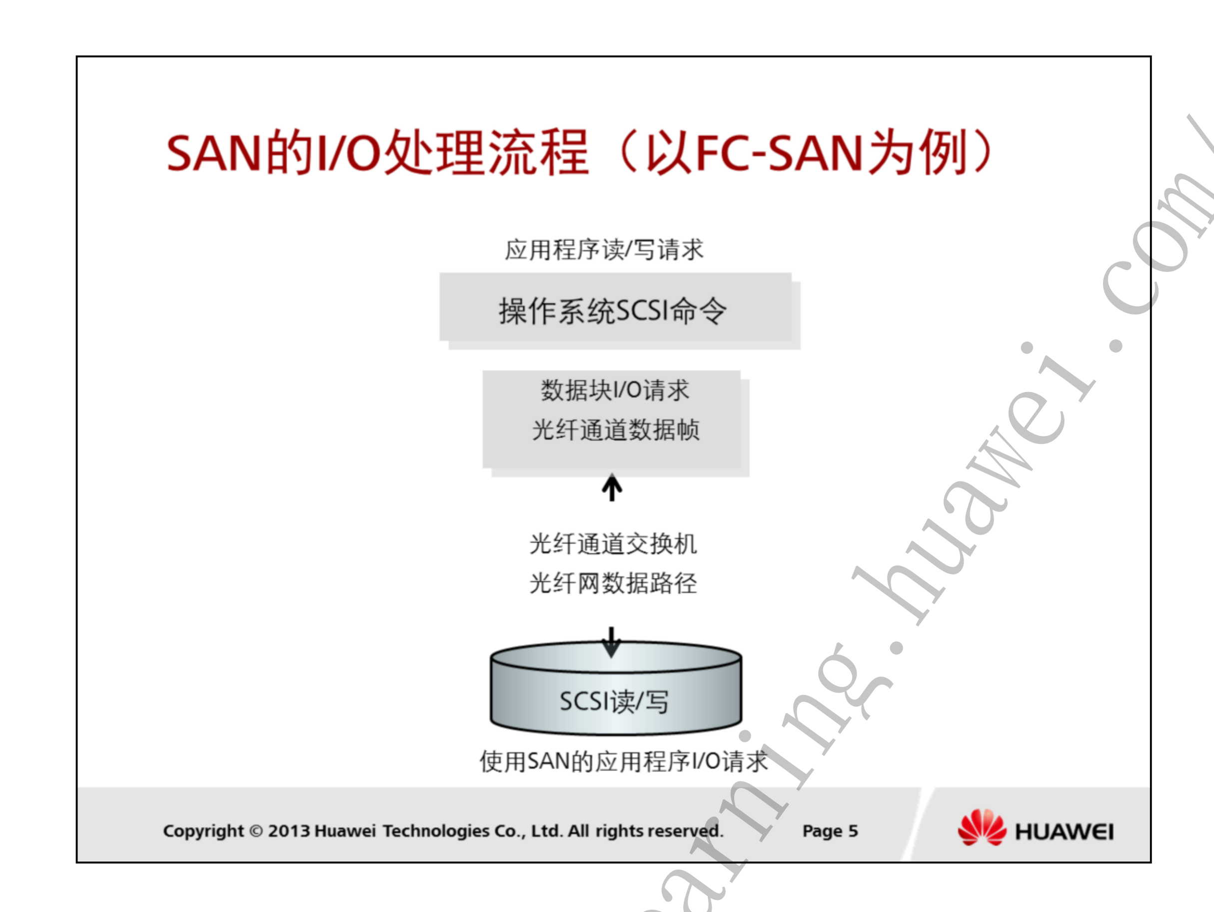


SAN即Storage Area Network（存储区域网络），使用专用网络连接主机和存储设备。目前使用的SAN网络协议主要有FC及iSCSI。



SAN的硬件组件主要有光纤交换机/以太网交换机、HBA卡/网卡、FC-SAN存储设备/IP-SAN存储设备。

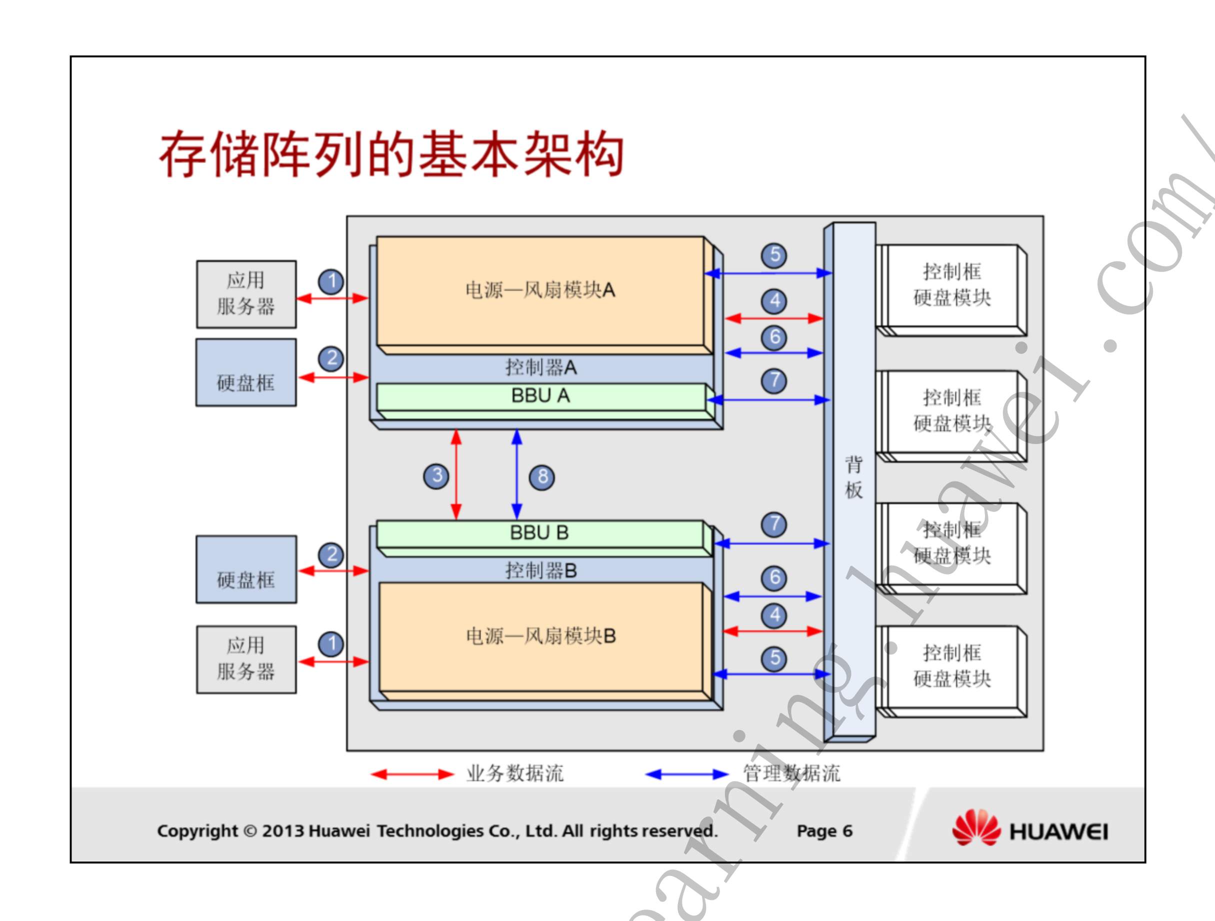
主机总线适配器（Host Bus Adapter，HBA）是连接服务器与存储区域网络的设备，包括FC-HBA卡、SAS-HBA卡和iSCSI HBA卡。



SAN提供了一种存储互联的机制，使得其上所连的服务器能够像与直接连接存储设备通信一样处理I/O请求。这使得网络上连接的服务器的操作系统可以通过存储网络中的光纤通道协议传输并执行数据块I/O。因此，SAN具有数据块I/O操作的种种优点和效率，同时也具有网络的灵活性和更高性能。

SAN的数据块I/O要求与发出请求的操作系统保持直接的通信。

应用程序的读写请求通过操作系统转化成SCSI请求，然后进一步封装成FC请求，随后该请求经过光纤网络数据路径下发到目标设备上（FC磁盘），目标设备执行完成该请求后又通过原来的路径将结果返回给应用程序。



Oceanstor 系列存储在两个层面上实现了对数据的冗余保护，第一个层面：Oceanstor 系列存储产品所有FRU器件均是冗余的，可以现场热插拔，没有单点故障；第二个层面：Oceanstor 系列存储实现了双控制器的双控双活保护，保证最重要的存储控制器没有单点故障。

上图中，数据流的关系如下：

① 控制器和应用服务器之间的业务数据流

② 控制器和硬盘框之间的业务数据流

③ 控制器A 和控制器B 之间的镜像业务数据流

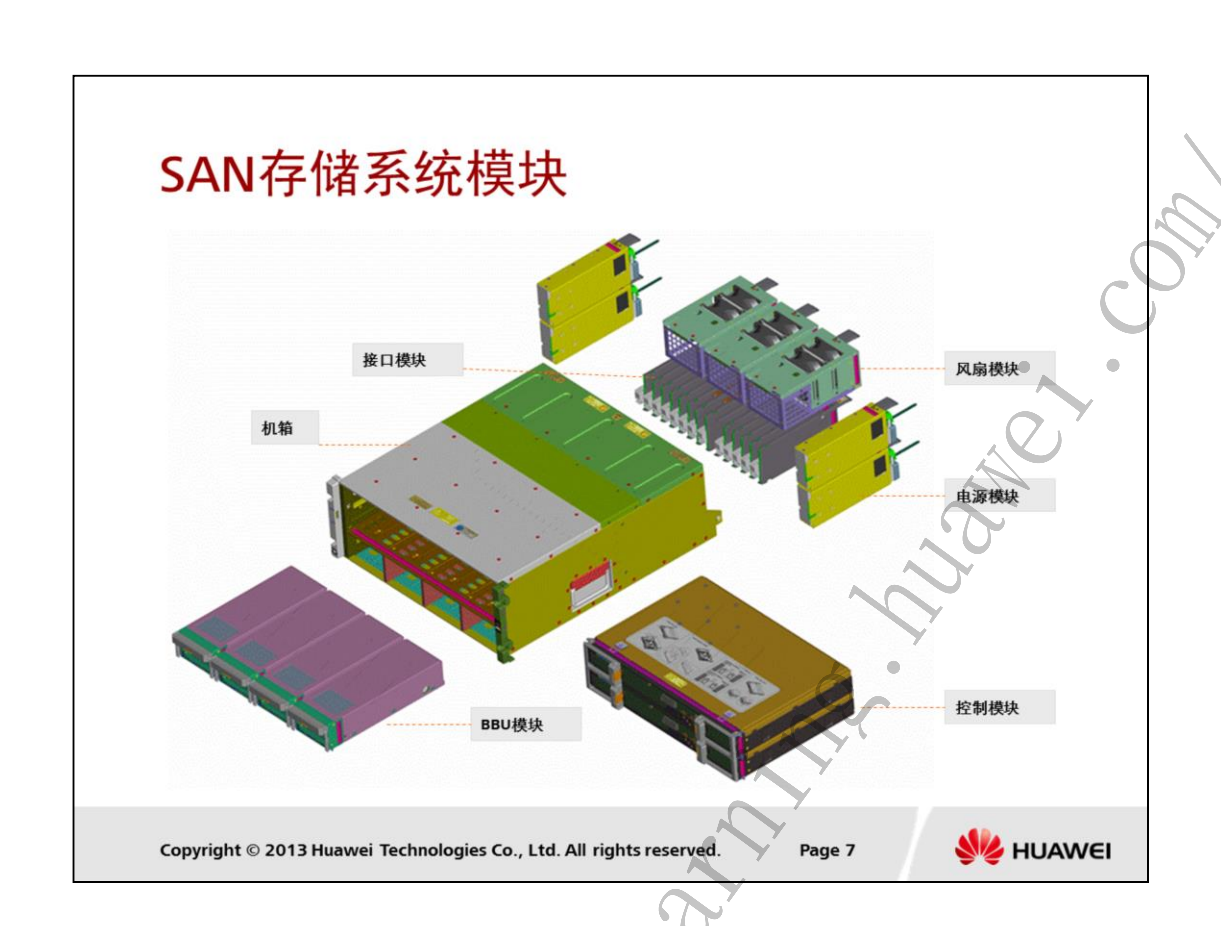
④ 控制器和硬盘模块之间的业务数据流

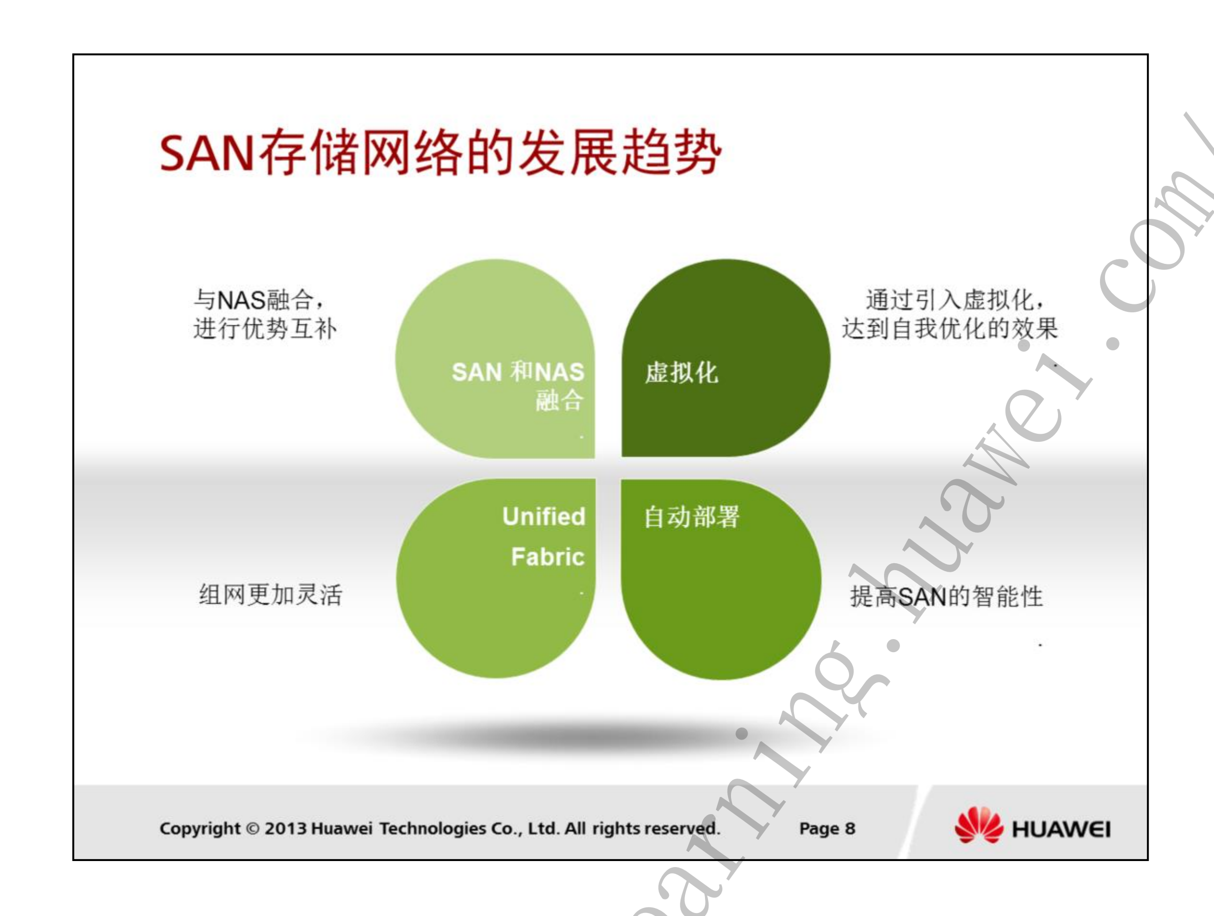
⑤ 控制器和电源—风扇模块之间的管理数据流

⑥ 控制器和硬盘模块之间的管理数据流

⑦ 控制器和BBU 之间的管理数据流

⑧ 控制器A 和控制器B 之间的心跳管理数据流


SAN存储系统模块
接口模块
风扇模块
机箱
电源模块
控制模块
BBU模块
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 7
HUAWEI



- SAN与NAS融合：SAN可扩展性好，性能高而NAS成本低管理方便，二者的有效结合最大程度满足客户需求。
- 虚拟化：资源整合、动态负载均衡是SAN发展的必然趋势，也是虚拟化的重要特点。
- Unified Fabric：目前有多种结构的SAN，如FC-SAN，IP-SAN，IB-SAN（即InfiniBand-SAN）等，渐渐出现了一些融合的趋势，如FCoE等，最终都走向了以太网。
- 智能性：自动部署，性能自我调优，存储不再绝对依赖于管理员的操作，避免人为错误也提高了性能。



目录



SCSI协议基础

- SCSI:小型计算机接口（Small Computer System Interface）
- SCSI协议本质上同传输介质无关，SCSI可以在多种介质上实现，甚至是虚拟介质；
- 目前最重要的存储协议；
- 目前已发展到SCSI-3。

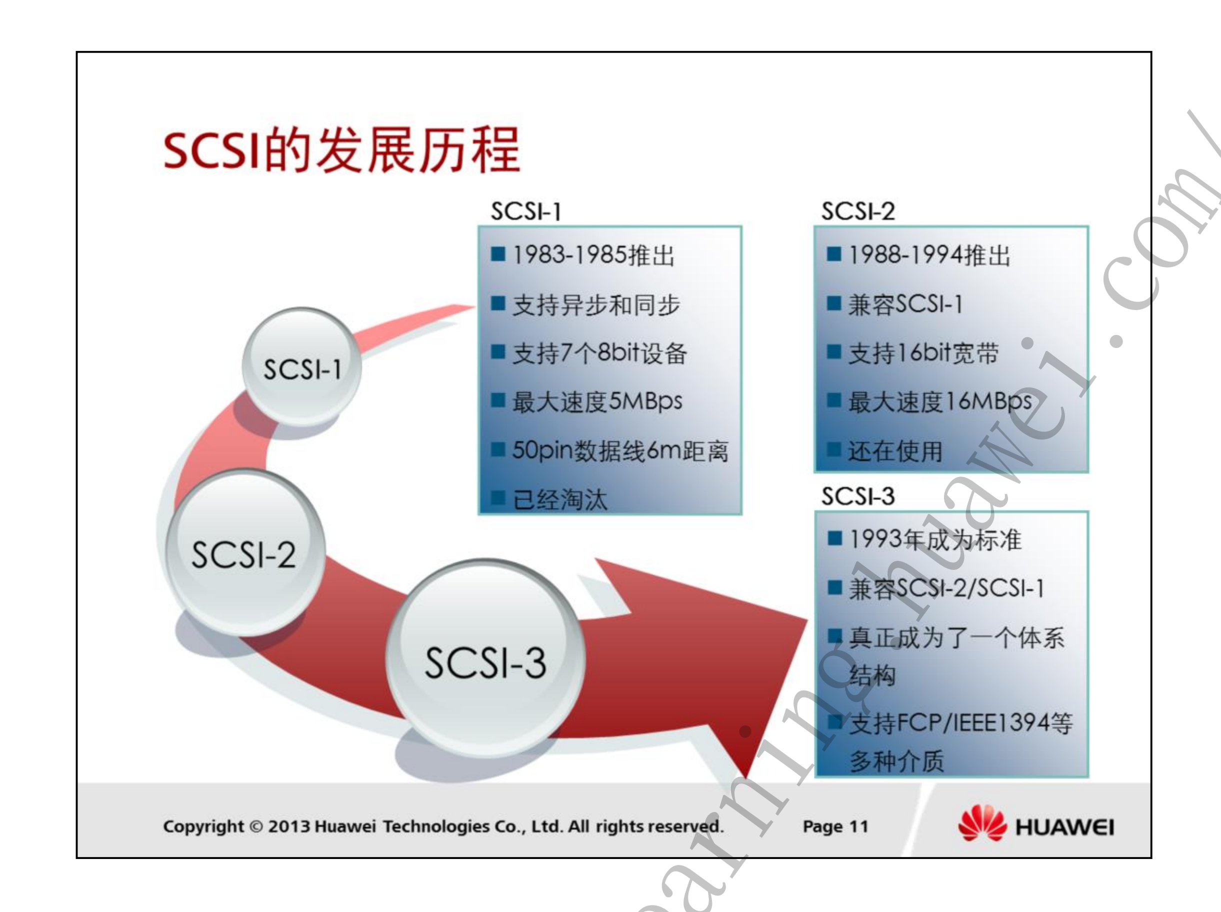
 Page 10 

SCSI指的是一个庞大协议体系，到目前为止经历了SCSI-1/SCSI-2/SCSI-3变迁。

SCSI协议定义了一套不同设备（磁盘，磁带，处理器，光设备，网络设备等）利用该框架进行信息交互的模型和必要指令集。

目前，SCSI协议可以在FC链路协议、基于SAS的链路协议、基于虚拟IP链路的iSCSI协议等多种介质上实现。

小型计算机系统接口（SCSI，Small Computer System Interface）是一种用于计算机及其周边设备之间（硬盘、软驱、光驱、打印机、扫描仪等）系统级接口的独立处理器标准。SCSI标准定义了命令、通信协议以及实体的电气特性（换成OSI的说法，就是占据了实体层、链接层、通信层、应用层），最大部份的应用是在存储设备上（例如硬盘、磁带机）；但其实SCSI可以连接的设备包括有扫描仪、光学设备（像CD、DVD）、打印机……等等，SCSI命令中有条列出支持的设备SCSI周边设备。



- SCSI-1

1983年开始研究、1986年制定的SCSI标准的主要特点是：支持同步和异步的SCSI设备；支持7台8位的SCSI设备；异步传输速率最大为1.5M/秒，同步传输输率最大为5M/秒；支持WORM（WRITE ONCE READ MANY）设备。SCSI－1控制使用卡ISA总线，它的最大连线长度为6米，接头为50针。但由于传输速度太慢，现在已经不使用了。

- SCSI-2

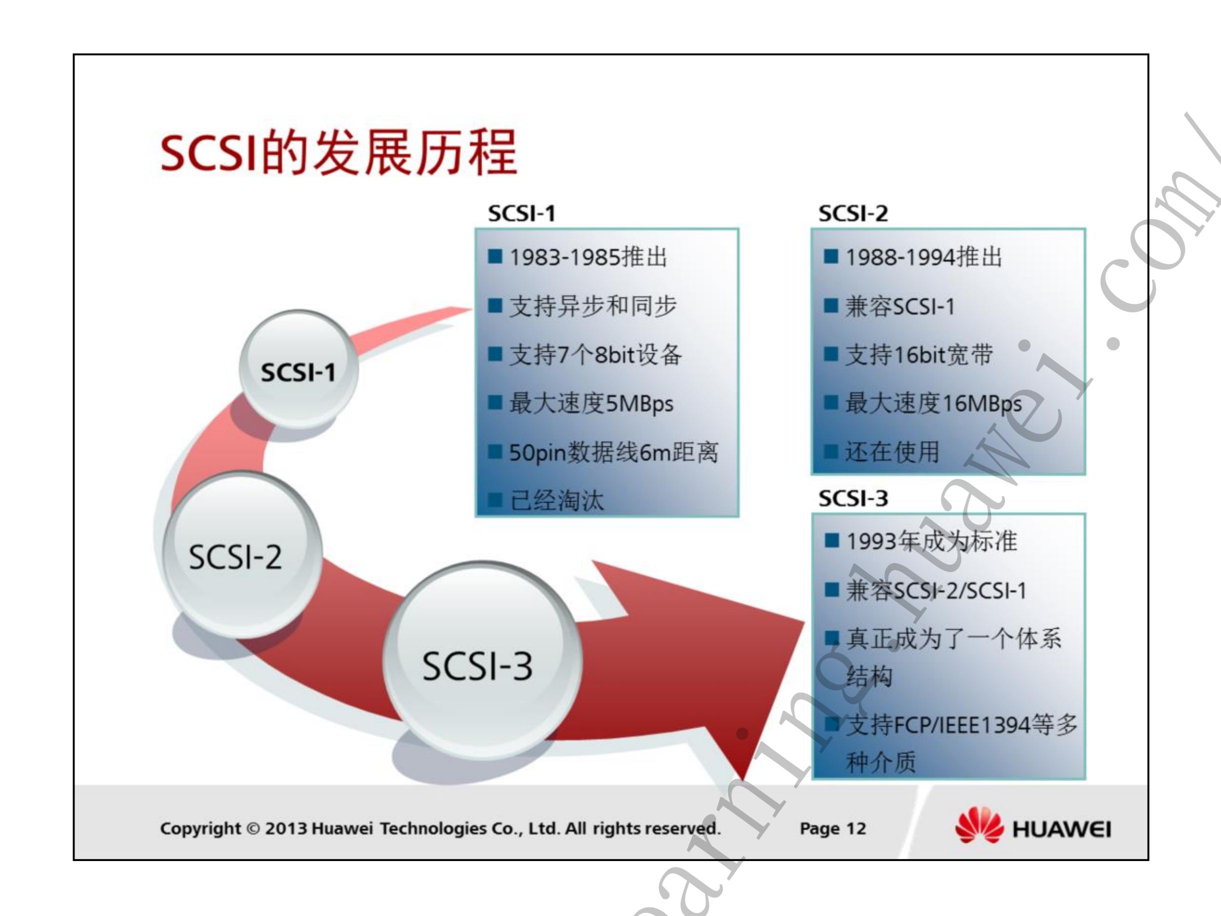
SCSI－2标准是1992年制定的，它在SCSI－1标准中加入以下新功能；支持高密度SCSI接头；支持CD－ROM和扫描仪； SCSI总线具有奇偶校验功能； 支持FAST SCSI 和WIDE SCSI； 支持Tagged Queuing功能。

- SCSI-3

与SCSI－2相比SCSI－3能支持更多的计算机硬件种类，并且数据传输率也更快。SCSI－3 支持Ultra SCSI，SCSI－3也叫FAST－20 、Doublespeed SCSI，它定义怎样在8位SCSI总线上每秒传输20M数据和在16位Wide SCSI总线上每秒传输40M数据。

- Ultra SCSI

在这以前的SCSI总线，只能在上升沿或下降沿传输数据，Ultra SCSI可在同时在上升沿和下降沿传输数据，这使得在16位数据总线上传输速率达到40M/s 。



- ULTRA 160 SCSI:

将并行SCSI总线的带宽提高一倍。使用16位数据总线，LVD技术、双边时钟传输技术，可以使同步传输带宽达到160MB/s 。使用了CRC提高了数据完整性，并根据误码率动态调整传输速率。

- ULTRA 320 SCSI

16位数据总线，使用LVD技术，相比ULTRA 160 SCSI，它时钟输率提高了一倍，使用快速总裁和选择技术、读写数据流技术。

SCSI-3接口规格：

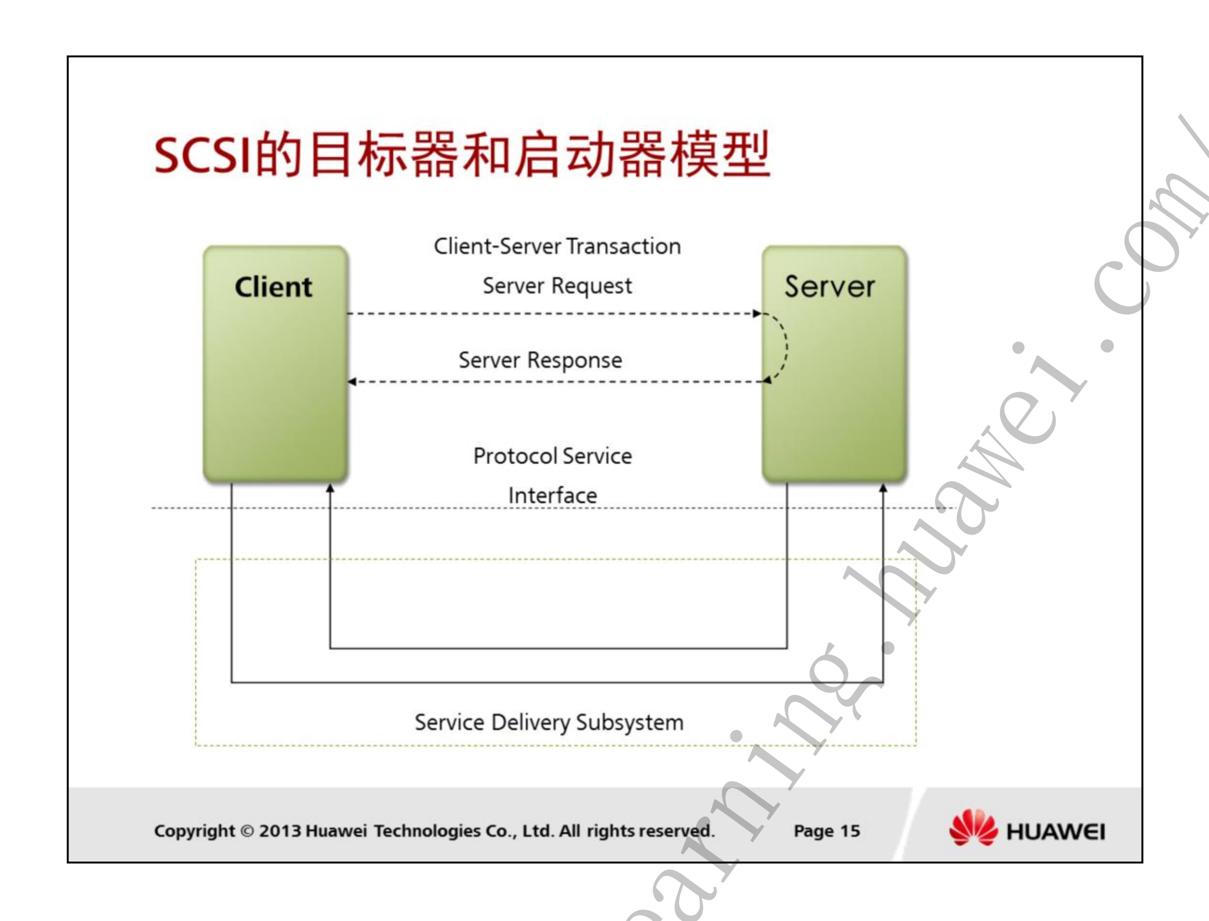
规格	别名	规范文件	接口	数据包 (bits)	频率	各种极限		
						带宽 (MB/s)	带宽 (Mbit/s)	设备数
Ultra SCSI	Fast-20	SCSI-3 SPI	IDC50	8	20 MHz	20 MB/s	160 Mbit/s	8-4 (HVD:8)
Ultra Wide SCSI		SCSI-3 SPI	68-pin	16	20 MHz	40 MB/s	320 Mbit/s	8-4 (HVD:16)
Ultra2 SCSI	Fast-40	SCSI-3 SPI-2 (1997)	50-pin	8	40 MHz	40 MB/s	320 Mbit/s	8
Ultra2 Wide SCSI		SCSI-3 SPI-2	68-pin; 80-pin (SCA/SCA-2)	16	40 MHz	80 MB/s	640 Mbit/s	16
Ultra3 SCSI	Ultra-160; Fast-80 wide	SCSI-3 SPI-3 (1999)	68-pin; 80-pin (SCA/SCA-2)	16	40 MHz DDR	160 MB/s	1280 Mbit/s	16
Ultra-320 SCSI	Ultra-4 SCSI or Fast-160 SCSI	SCSI-3(2002)	68-pin; 80-pin (SCA/SCA-2)	16	80 MHz DDR	320 MB/s	2560 Mbit/s	16
Ultra-640 SCSI	Ultra-5;	SCSI-3(2003)	68-pin; 80-pin	16	160 MHz DDR	640 MB/s	5120 Mbit/s	16



Interface	Alternative names	Width (bits)	Clock	Maximum		
				Throughput (MB/s)	Throughput (Mbit/s)	Devices
SSA	Serial Storage Architecture	1	200 MHz	40 MB/s	320 Mbit/s	96
SSA 40		1	400 MHz	80 MB/s	640 Mbit/s	96
Fibre Channel 1Gbit	1GFC	1	1 GHz	100 MB/s	800 Mbit/s	127 (FC-AL)/2^{24} (FC-SW)
Fibre Channel 2Gbit	2GFC	1	2 GHz	200 MB/s	1600 Mbit/s	127/2^{24}
Fibre Channel 4Gbit	4GFC	1	4 GHz	400 MB/s	3200 Mbit/s	127/2^{24}
Fibre Channel 8Gbit	8GFC	1	8 GHz	800 MB/s	6400 Mbit/s	127/2^{24}
Fibre Channel 16Gbit	16GFC	1	16 GHz	1600 MB/s	12.8 Gbit/s	127/2^{24}
SAS 1.1	Serial attached SCSI	1	3 GHz	300 MB/s	2400 Mbit/s	16,256
SAS 2.1		1	6 GHz	600 MB/s	4800 Mbit/s	16,256
SAS 3.0		1	12 GHz	1200 MB/s	9600 Mbit/s	16,256
IEEE 1394	Serial Bus Protocol (SBP)	1		400 MB/s	3200 Mbit/s	63
SCSI Express	SCSI over PCIe (SOP)	1	8 GT/s	985 MB/s	7877 Mbit/s	
USB Attached SCSI	UAS	1	5 Gbit/s	~400 MB/s	~3200 Mbit/s	127
ATAP	ATA Packet Interface	16	33 MHz DDR	133 MB/s		2
iSCSI			implementation- and network-dependent			2^{128} (IPv6)
SRP	SCSI RDMA Protocol (SCSI over InfiniBand and similar)		implementation- and network-dependent			



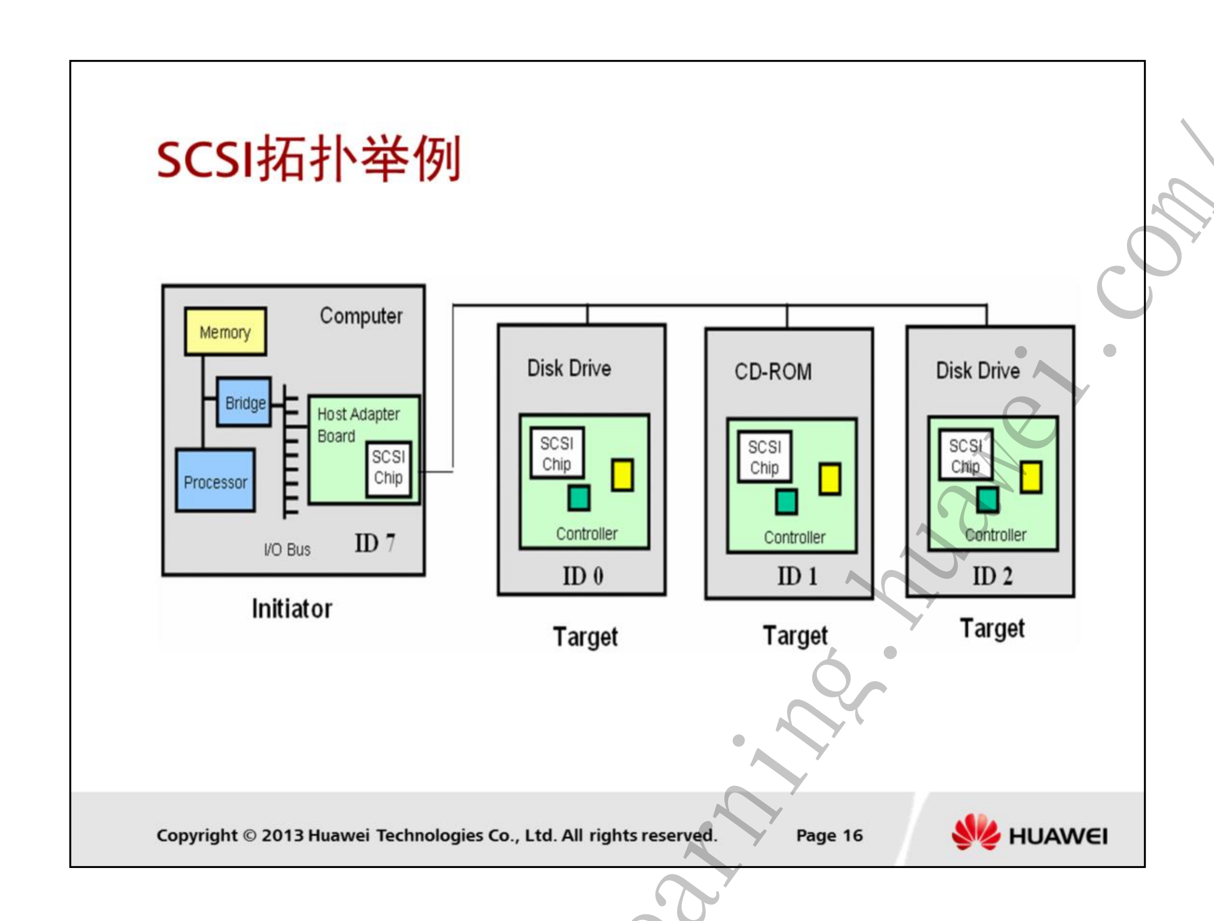
- 其他相关接口介绍



Client, 即initiator端，向Server发起请求，Server完成请求并给出response。

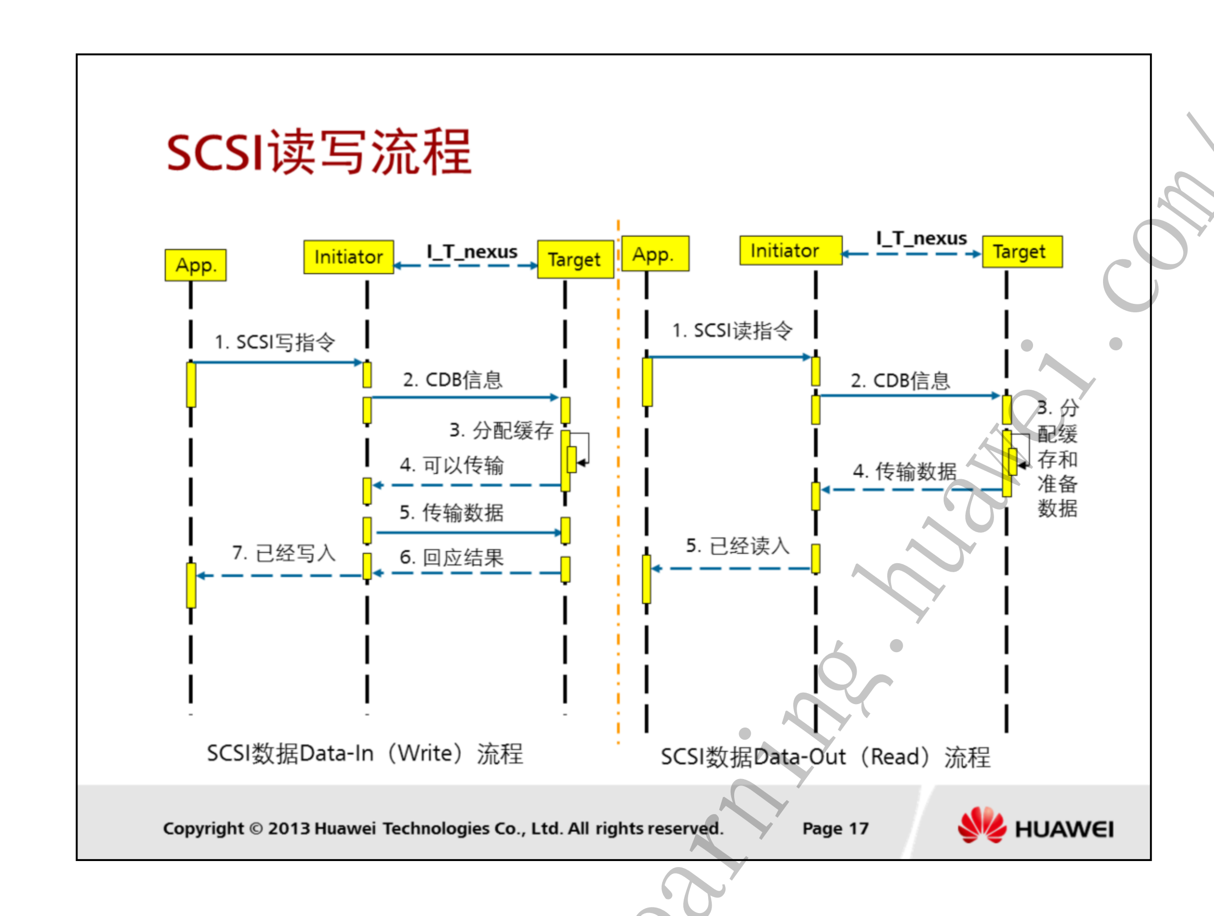
Server,即target端，接受client发过来的请求并执行，最后将结果返回给client。

Initiator和target之间传输所使用的服务传输子系统（Service Delivery SubSystem）可以使用SAS协议、FC协议等。



SCSI是典型的I-T模式，initiator（发起端）通过线缆与target（目标端）连接，initiator负责发起命令，而target端负责解析命令，处理命令并向initiator返回结果。

比较典型的SCSI initiator有iSCSI initiator、FC HBA、SAS HBA等，典型的SCSI target有磁盘、CD-ROM等。



- SCSI写流程：

  1、应用客户端发起SCSI写请求。

  2、SCSI Initiator收到该请求，将该命令发给SCSI Target。

  3、SCSI Target 收到写请求，分配本地缓存（用于存放写请求的数据）。

  4、SCSI Target 通知SCSI Initiator本地缓存分配完成，可以进行数据传输。

  5、SCSI Initiator向SCSI Target发送数据。

  6、SCSI Target 向SCSI Initiator回应写的结果。

  7、SCSI Initiator收到该结果并将其通知应用程序，至此，写流程完成。

- SCSI读流程：

  1、应用客户端发起SCSI 读请求。

  2、SCSI Initiator收到该请求，将该命令发给SCSI Target。

  3、SCSI Target 收到读请求，分配本地缓存（用于准备读请求的数据）。

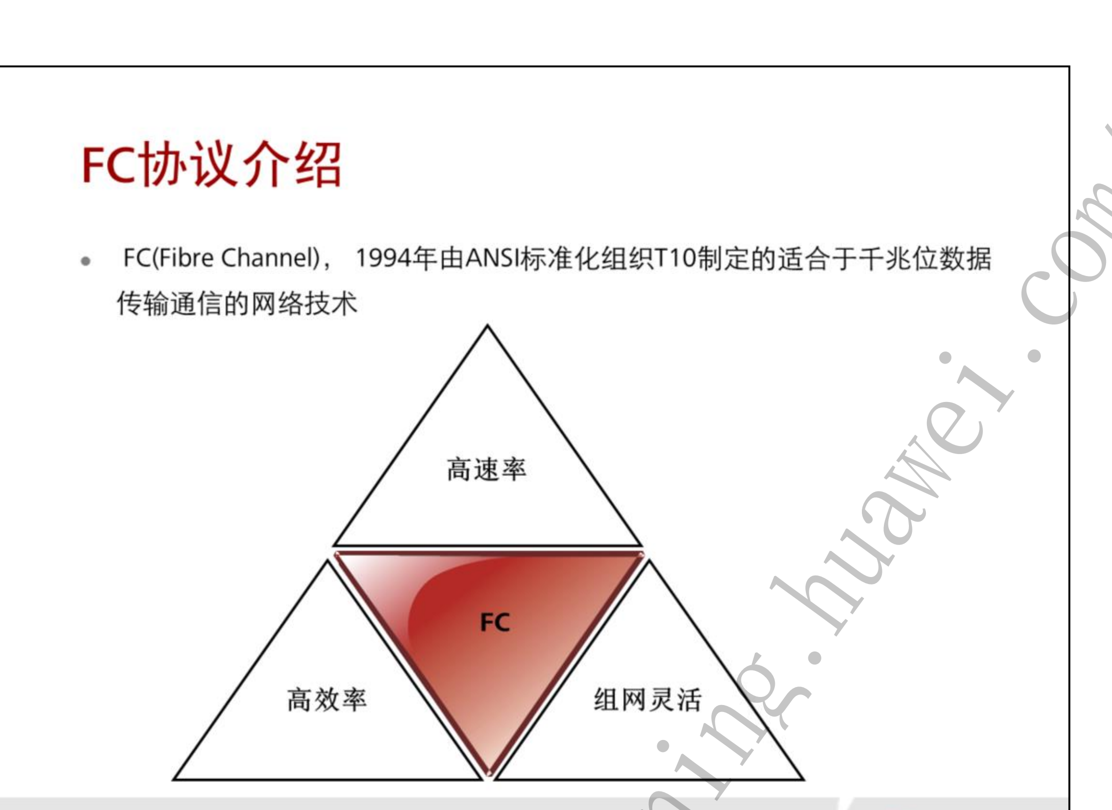
  4、SCSI Target将数据传输给SCSI Initiator。

  5、SCSI Initiator通过应用程序读操作完成。

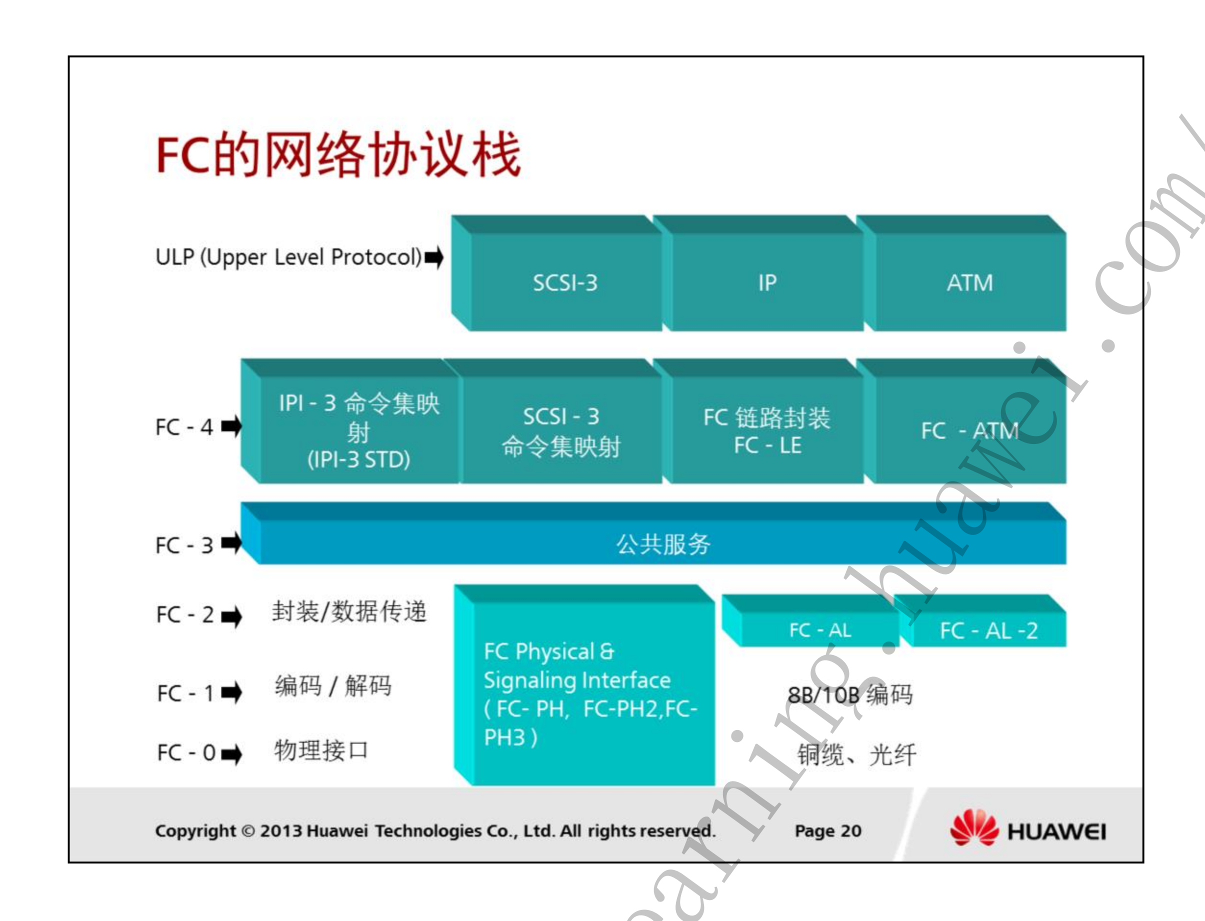


目录





- 高速率：目前单个FC端口的速率可以达到16Gb/s。
- 高效率：FC 极低的误码率（10的负12次方）、可靠的硬件、8B/10B编码、灵活的底层错误恢复机制等使其链路利用率非常高。
- 组网灵活：支持Fabric、Loop、P2P等组网方式并支持这几种方式的混合组网。



- FC网络协议栈共分5层，分别为FC-0，FC-1，FC-2，FC-3，FC-4。
  - FC-0：连接物理介质的界面、电缆等；定义编码和解码的标准。
  - FC-1：传输协议层或数据链接层，编码或解码信号。
  - FC-2：网络层，光纤通道的核心, 定义了帧、流控制、和服务质量等。
  - FC-3：定义了常用服务，如数据加密和压缩。
  - FC-4：协议映射层，定义了光纤通道和上层应用之间的接口，上层应用比如：串行SCSI 协 议，HBA卡的驱动提供了FC-4 的接口函数。FC-4 支持多协议，如：FCP-SCSI，FC-IP，FC-VI。
  - 光纤通道的主要部分实际上是FC-2。其中从FC-0到FC-2被称为FC-PH，也就是“物理层”。光纤通道主要通过FC-2来进行传输，因此，光纤通道也常被成为“二层协议”或者“类以太网协议”。

光纤网的拓扑

点对点

FC-P2P

只能连接 2个设备

(直接连接)

光纤仲裁环

FC-AL

最多支持126个设备

(光纤集线器)

交换式光纤网

FC-SW

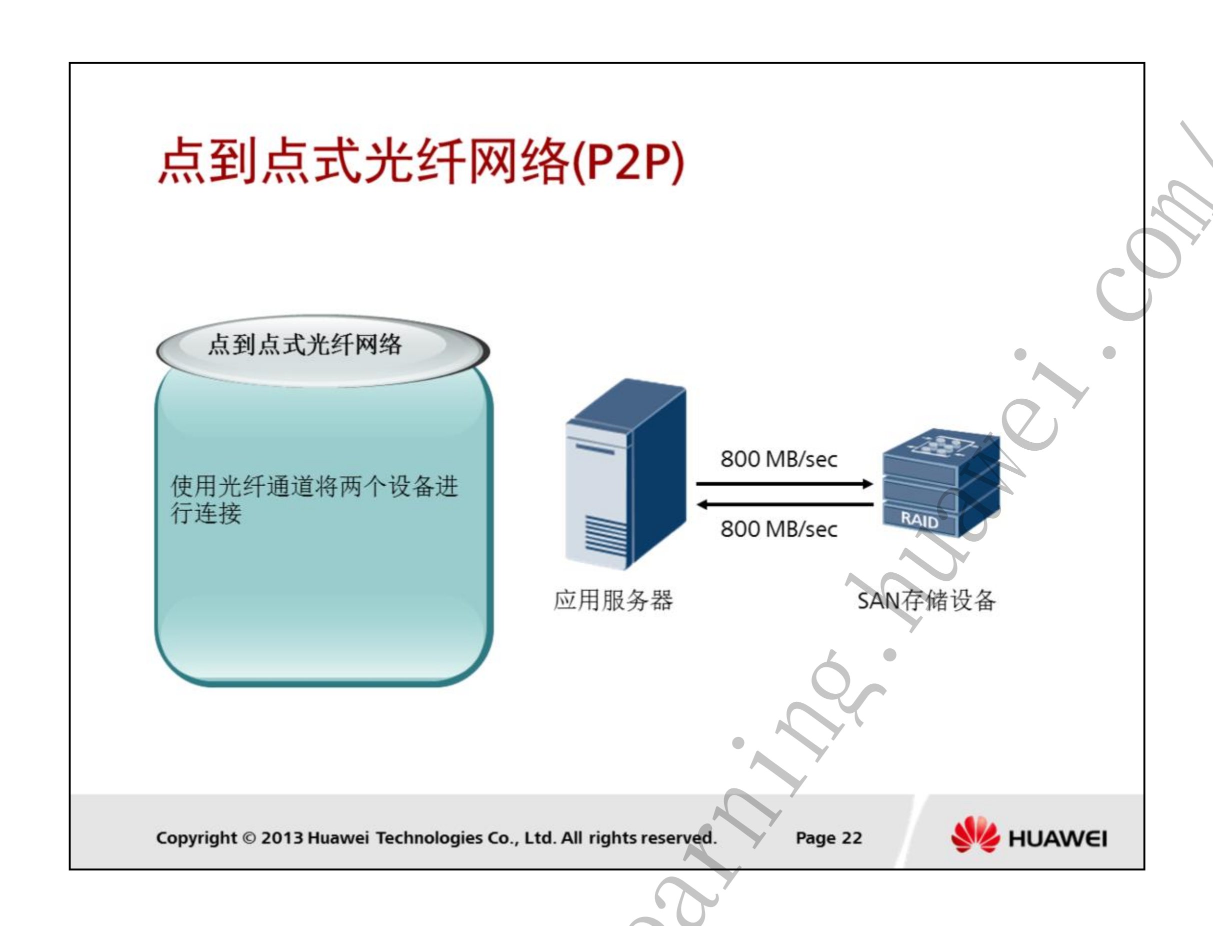
最多支持1600万个设备

(光纤网交换机)

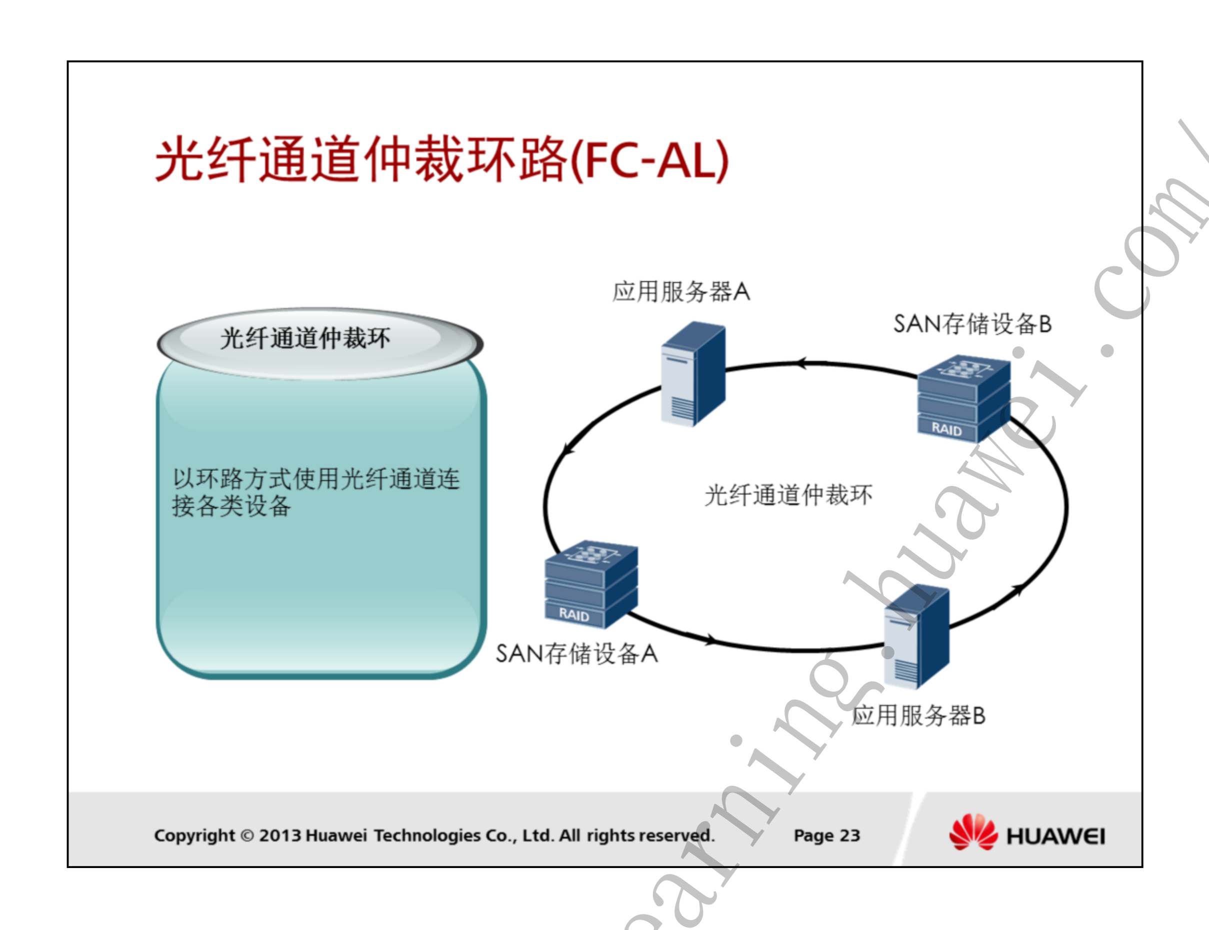
Page 21

HUAWEI

- PTP（点对点）：一般用于DAS（直连式存储）设置
  - 服务器和存储设备在点对点的环境里都是N_PORT. 通过一条上行一条下行两条通道进行数据存储与读取。
- FC-AL（光纤通道仲裁环路）：采用FC-AL仲裁环机制，使用Token（令牌）的方式进行仲裁。光纤环路端口，或交换机上的FL端口，和HBA上的NL端口（节点环）连接，支持环路运行。采用FC-AL架构，当一个设备加入FC-AL的时候，或出现任何错误或需要重新设置的时候，环路就必须重新初始化。在这个过程中，所有的通信都必须暂时中止。由于其寻址机制，FC-AL理论上被限制在了127个节点。
- FC-SW（FC Switched 交换式光纤通道）：在交换式SAN上运行的方式。FC-SW可以按照任意方式进行连接，规避了仲裁环的诸多弊端，但需要购买支持交换架构的交换模块或FC交换机。

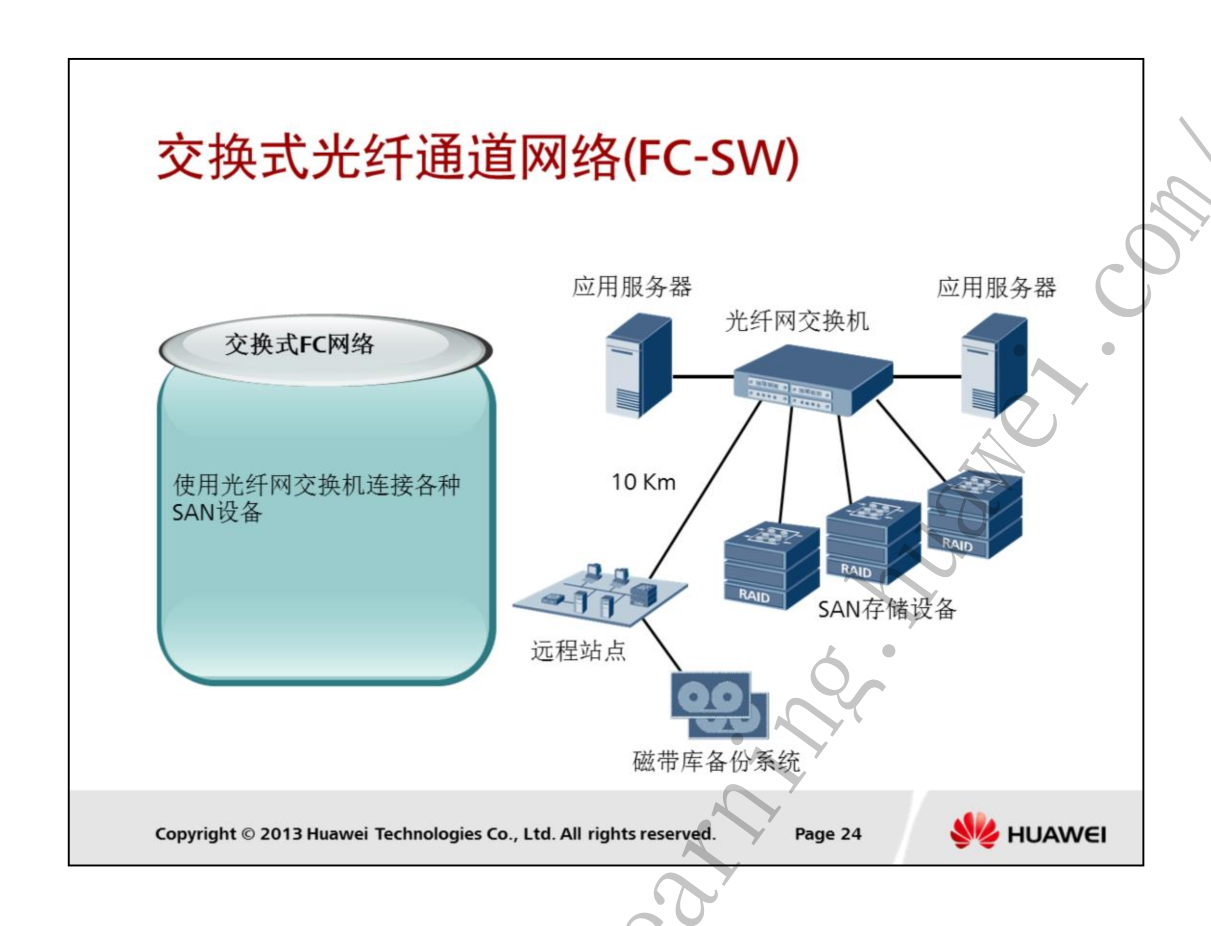


在点到点式光纤通道网络中，使用光纤通道将两个设备进行连接。这种方式用于光纤通道磁盘阵列的最初实施中，相对而言，它仍然是一种直接连接存储的解决方案，但是由于使用光纤网络而使得带宽得到了较大的提高。同时，长达10公里的传输能力远远超过了一般总线的能力。



光纤通道仲裁环以环路方式使用光纤通道对各类设备进行连接。仲裁环中的设备以环路方式配置，不但共享带宽，而且还必须对网络的通信进行仲裁。这种方式借助光纤通道网络的高带宽特性提升了速度，并且可以在网络中放置额外的磁盘阵列，并允许更多的服务器与之相连。光纤通道仲裁环结构中存在着较为严重的延迟问题，特别是随着加入到仲裁环中的设备数量增多，由于环路中所有设备共享带宽，所以每个设备实际能获取的带宽减少使得延迟增加。这也导致光纤通道仲裁环不可能达到交换式光纤通道光纤网络的性能水平。

仲裁环拓扑结构下，当有新的节点加入或者旧的节点退出时，环路需要重新协商。



交换式光纤通道网络是采用光纤网交换机构建的一种存储网络。这种网络交换机类似于数据网络中所使用的以太网交换机，不过它是基于FC协议而不是TCP/IP协议。交换式FC-SAN在建立点对点通信方面，类似于点对点网络。不同之处在于，它所建立的点对点连接并不是永久性的，可同时并发很多点对点的虚连接。通过光纤网络交换机连接起来的设备可以独享带宽，使得数据传送效率不再受到设备连接数量的影响。

FC的三种拓扑比较

特性	点到点式	仲裁环方式	光纤交换式
最大节点数	2	127	≈1600万
地址位数	无	8位AL_PA	24位端口地址
单点故障影响	无	环失效	无
多速率传输支持	否	否	是
数据帧传输顺序	按发送顺序	按发送顺序	无保证
介质访问方式	独享	仲裁式	独享

 Page 25



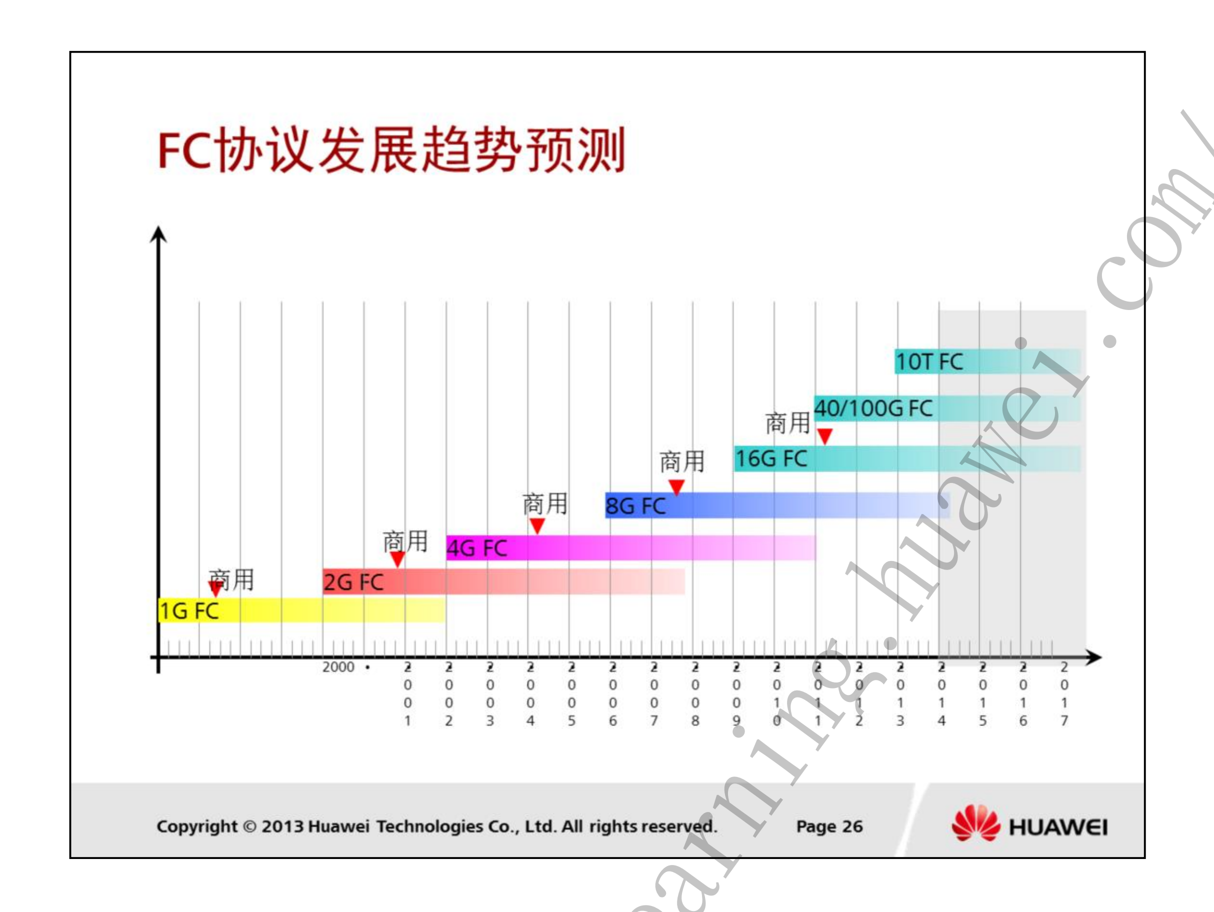
光线通道术语中的“节点”是指通过网络进行通信的任何实体，而不一定是一个硬件节点。这个节点通常是一个设备，比如说一个磁盘存储器、服务器上的一个主机总线适配器或者是一个光纤网交换机。

点到点式：两个设备背对背直接连接。这是最简单的一种拓扑，连接能力受限。

仲裁环式：这种设计方式中，所有设备连接在一个类似于令牌环的环路上。这个环路中添加或者移除一个设备会导致环路上所有活动中断。一个设备的故障导致整个环路不能进行工作。光纤通道集线器能够用于将众多设备连接到一起形成一个逻辑上的环路，并且能够旁路故障节点，使得环上节点的故障不会影响整个环路的通信。一个环路也可以通过使用线缆直接将节点一个接一个的连接成一个环而实现。最小的环路只包含两个节点，这种结构看起来和点到点式连接近似，它们的区别在很大程度上取决于各自的协议。

光纤交换式：所有的设备或者设备环都被连接到光纤网交换机上，与现有的以太网的实现形式在概念上是类似的。这种拓扑结构相对于点到点和仲裁环的优势在于：

- 交换机对结构形式进行管理，提供了最好的互联形式。
- 多对节点可以同时通信。
- 各个节点的故障是独立的，不会危及其他节点的正常工作。

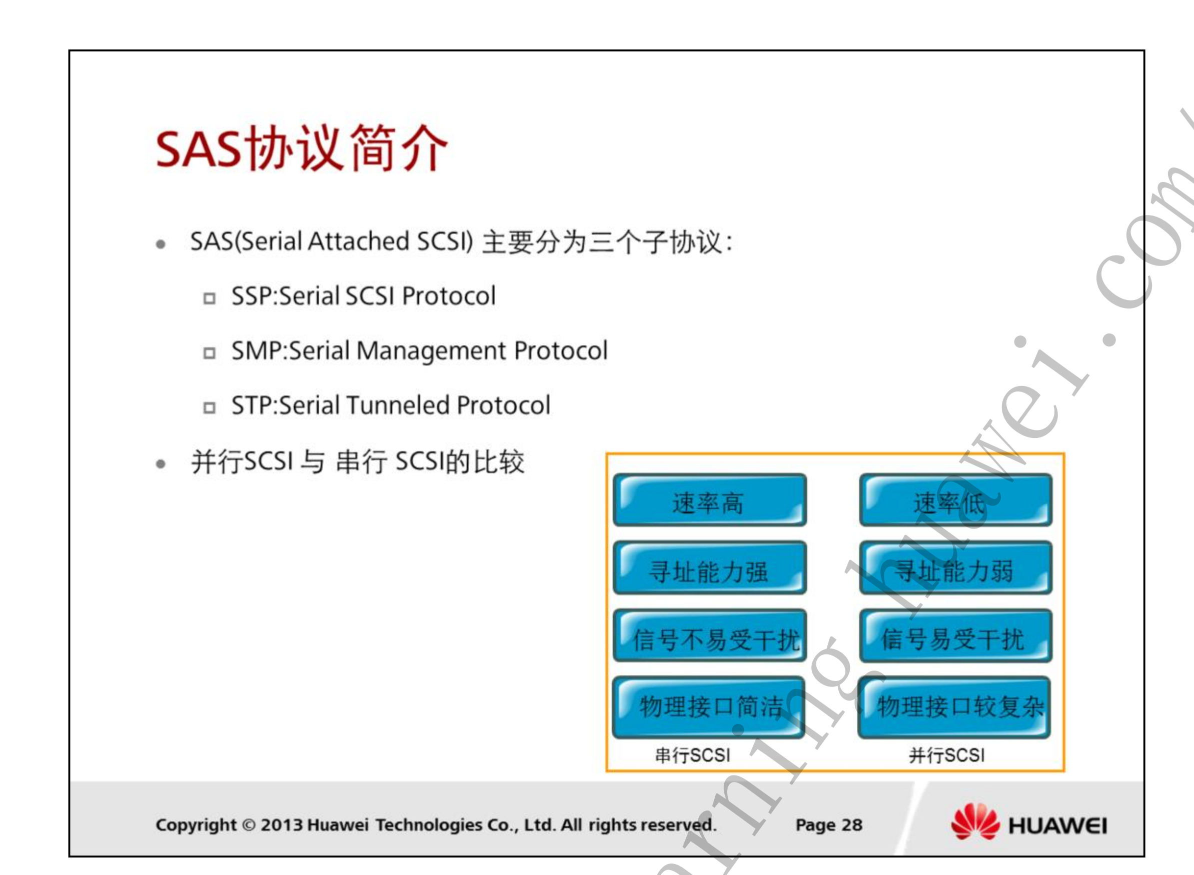


- 标准更新和商用规律：FC标准接口速率每隔36～48年个月进行一次更新，在新标准出来后，12～24个月实现商用(芯片厂商成熟推出芯片并有厂商开始采用)，主流厂商会在2年内陆续推出终端存储产品.。
- 每代标准商用生命周期在6～7年左右



目录

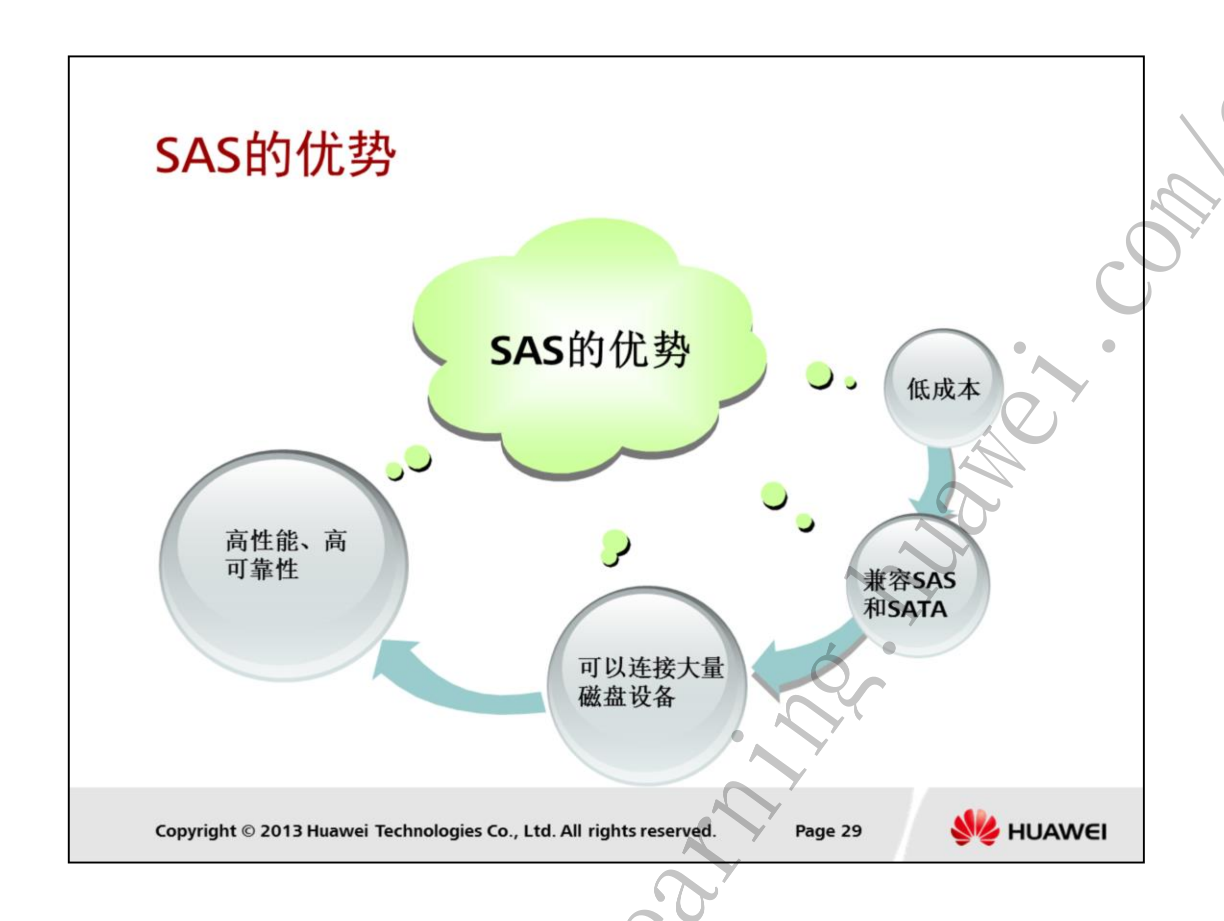




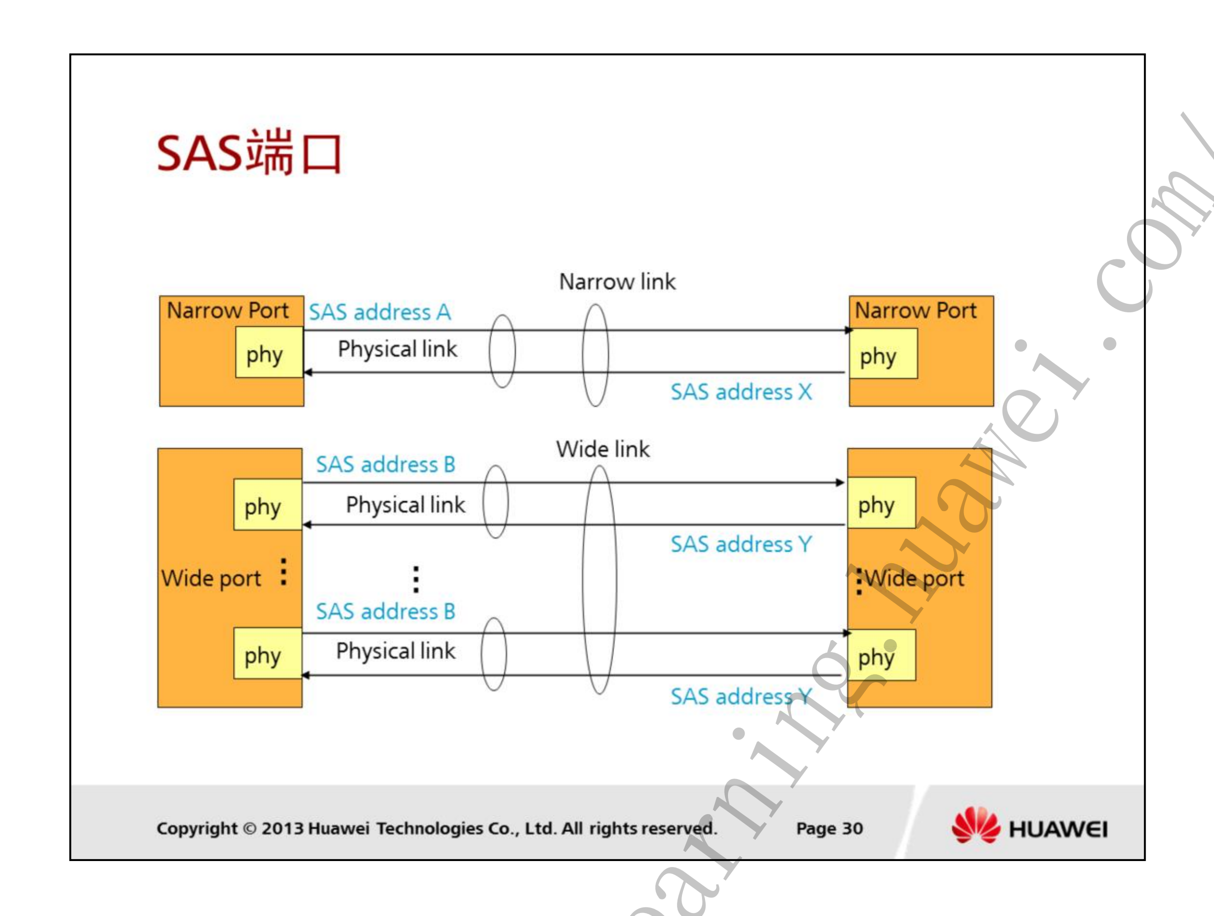
SAS 即串行连接SCSI，定义了一个新的串行点对点的企业级存储设备接口。SAS 1.0规范2003年9月由ANSI通过发布；SAS 2.0规范于2008年9月发布。

SAS协议主要包含以下三个子协议：

- SMP：SAS设备之间的信息。
- SSP：SAS 域中兼容SCSI命令。
- STP：SAS 域中兼容SATA命令。



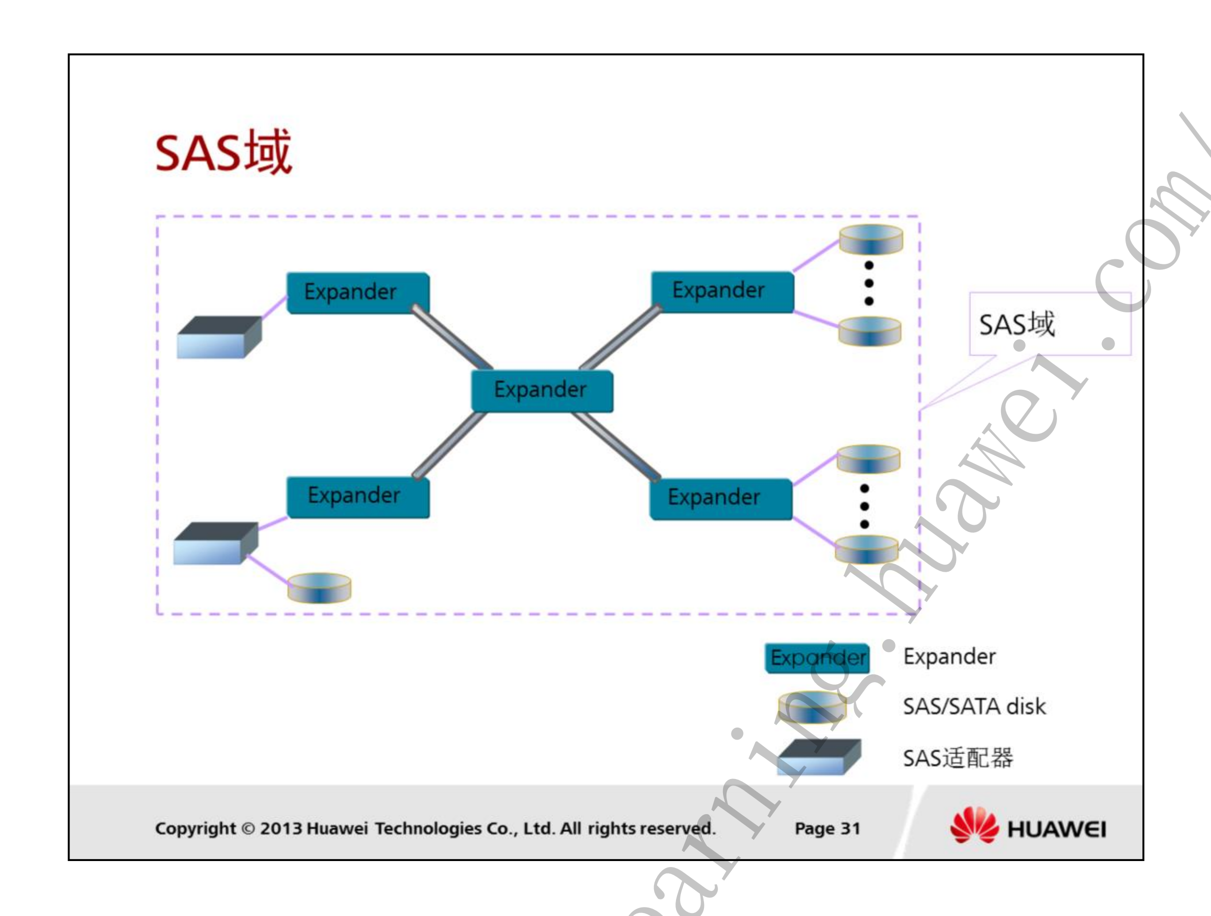
- 低成本：SAS相对于FC成本低很多。
- 兼容SAS、SATA：SAS的STP协议就是用来兼容SATA的。
- 连接设备多：典型的SAS组网最多可以容纳一万六千多个设备。
- 高性能：目前每个SAS Phy的速率可以达到6Gb/s,一般四个Phy组成一个端口，这样每个SAS端口的理论速率可以达到24Gb/s。



- Phy：SAS协议收发单元由一个发送器和一个接收器组成。
- 端口：一个或多个phy组成一个端口，一个phy组成的端口称为窄端口 （Narrow Port ），多个phy组成的端口称为宽端口（Wide port）。

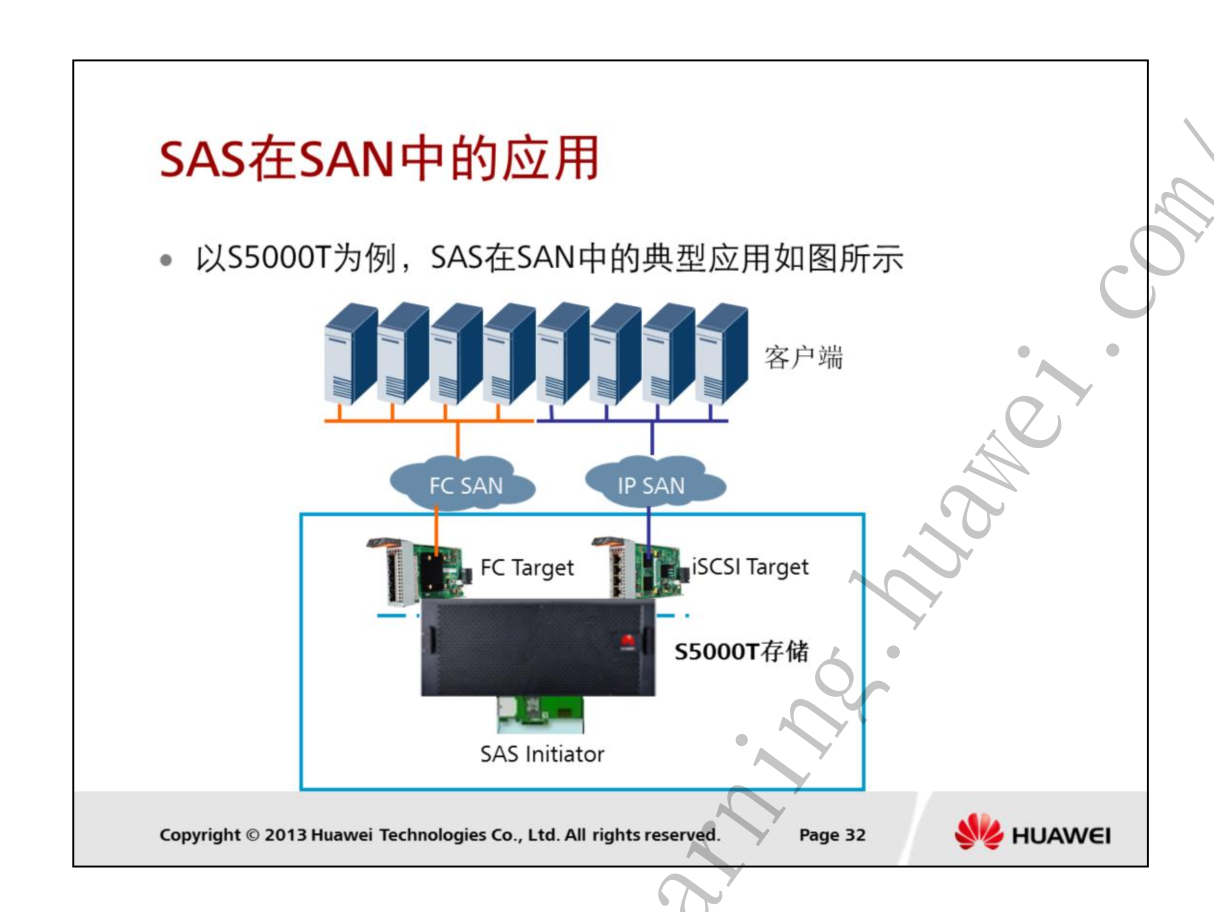
组成宽端口的几个phy其SAS地址相同，与之相连的端口的几个phy的SAS地址也必须相同。

以存储常用的miniSAS连接为例，由4个速率为6Gb/s的单端口组成宽端口，速率为24Gb/s



SAS域由以下几个部分组成：SAS Expander、终端设备、连接设备（即SAS连接线缆）。

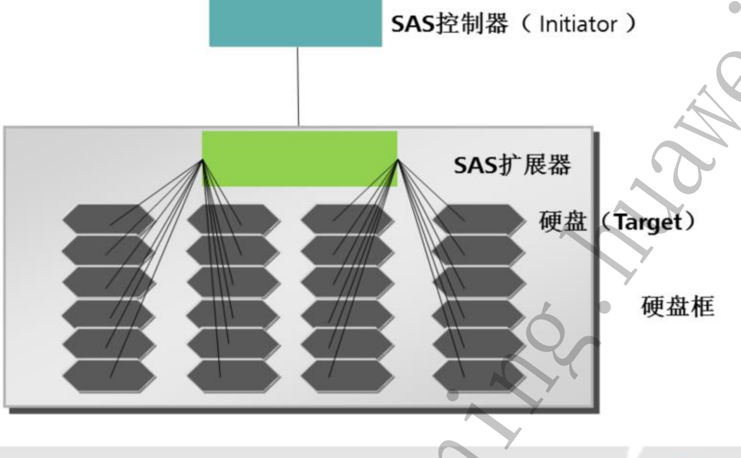
- SAS Expander：SAS域中的互联设备，类似于以太网交换机，通过 Expander的级联可以大大增加终端设备的连接数，从而节约HBA花费。终端设备主要有两种：
  - 启动器（通常为SAS HBA卡）。
  - 目标器（SAS/SATA硬盘，也可以是处于目标模式的HBA卡）。
- SAS 域中不能形成环路，以保证其discovery的正常进行。



应用客户端通过FC SAN或者IP SAN连接到S5000T的前端接口卡上（FC接口卡或者iSCSI接口卡），此时S5000T的前端接口卡作为目标器，应用客户端作为启动器。他们之间通过SCSI协议进行交互，而S5000T又通过其内置的SAS Initiator连接到磁盘框上，这些磁盘框是最终提供存储服务的物理存储设备。

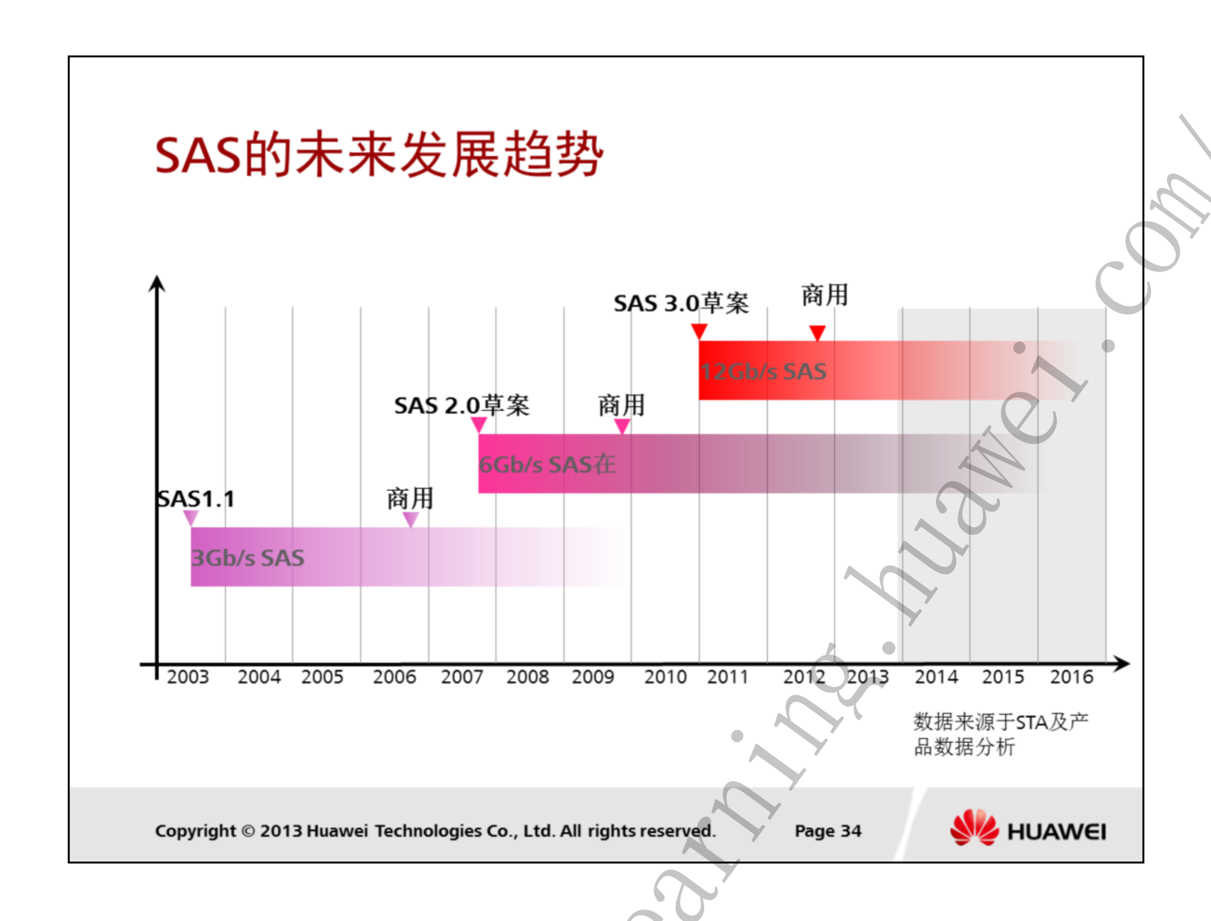
SAS在SAN中的应用

- 在阵列内部，所有硬盘均为Target，存储控制器作为Initiator对所有硬盘进行控制。



Page 33

HUAWEI



标准更新和商用规律：标准每隔36～40个月更新一次，传输速率翻番(SAS 1.1由于起步推广期较长).下一次更新周期预计在2011年下半年．由于IC厂商争相抢占先机，在新标准草案发布后12 ～18个月左右开始商用(首先是服务器HBA卡市场)，存储终端厂商在标准草案发布后24个月以后开始应用于终端产品。.

每一代标准商用生命周期在 4.5～5年左右。目前主流产品已经支持6GSAS接口。.

- 6G SAS：

  2010年，华为推出S5500T/S5600T/S6800T。

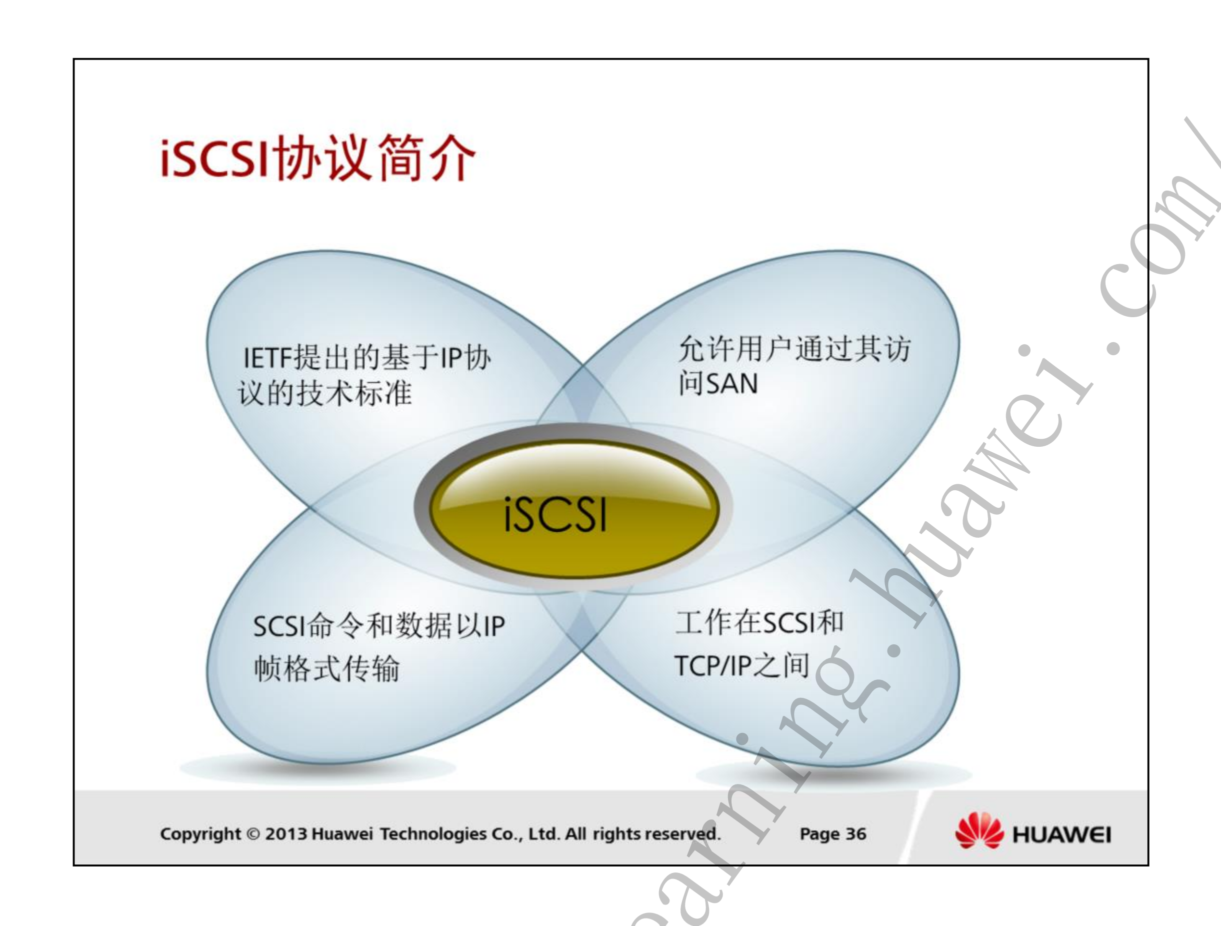
- 12G SAS：

  2013年，华为S5500T/S5600T/S6800T支持12GSAS协议。



目录

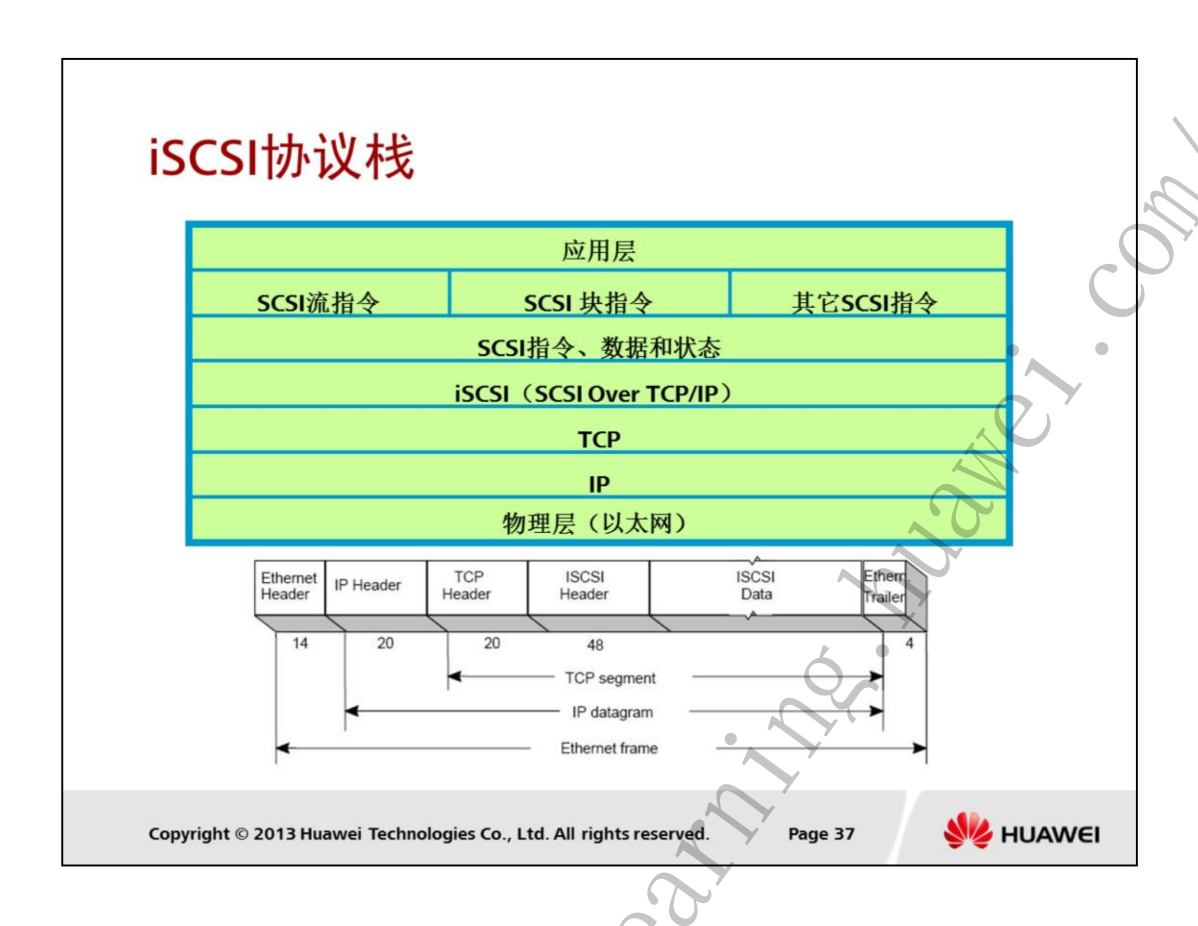




通过SCSI控制卡的使用可以连接多个设备，形成自己的“网络”，但是这个“网络”仅局限于与所附加的主机进行通信，并不能在以太网上共享。那么，如果能够通过SCSI协议组成网络，并且能够直接挂载到以太网上，作为网络节点和其它设备进行互联共享，那么SCSI就可以得到更为广泛的应用。所以，经过对SCSI的改进，就推出了iSCSI这个协议。基于iSCSI协议的IP-SAN是把用户的请求转换成SCSI代码，并将数据封装进IP包内在以太网中进行传输。

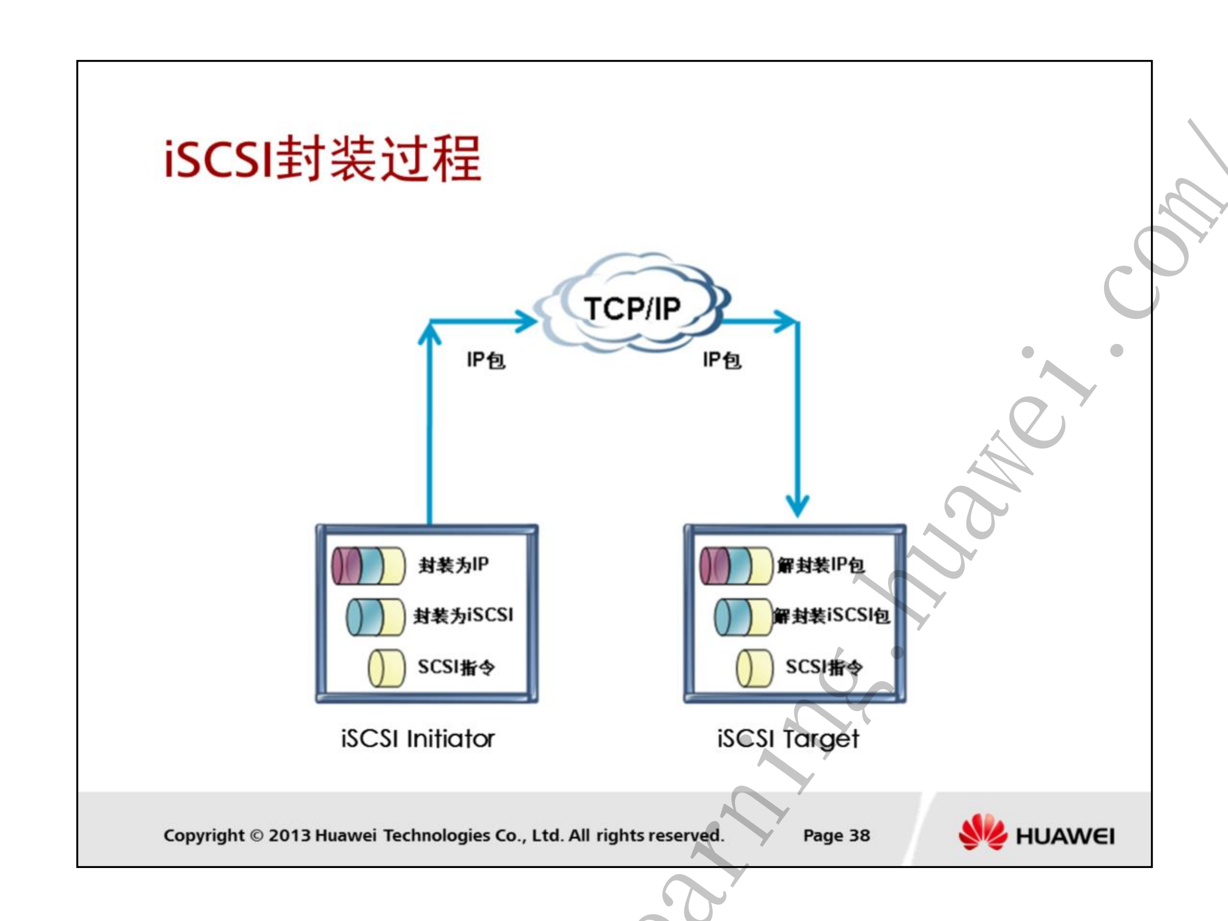
iSCSI方案最早是由Cisco和IBM两家发起，并且由Adaptec、Cisco、HP、IBM、Quantum等公司共同倡导。它提供基于TCP传输，将数据驻留与SCSI设备的方法。iSCSI标准草案在2001年推出，并经过多次论证和修改，于2002年提交IETF，在2003年2月，iSCSI标准正式发布。iSCSI技术的重要贡献在于其对传统技术的继承和发展：其一，SCSI（Small Computer System Interface，小型计算机系统接口）技术是被磁盘、磁带等设备广泛采用的存储标准，从1986年诞生起到现在仍然保持着良好的发展势头；其二，沿用TCP/IP协议，TCP/IP在网络方面是最通用、最成熟的协议且IP网络的基础建设非常完善。这两点为iSCSI的无限扩展提供了坚实的基础。

IP网络的普及性将使得数据可以通过LAN、WAN或者是通过Internet利用新型IP存储协议传输，iSCSI既是在这个思想的指导下进行研究和开发的。iSCSI是基于IP协议的技术标准，实现了SCSI和TCP/IP协议的融合，对众多的以太网用户而言，只需要极少的投资，就可以方便、快捷地对信息和数据进行交互式传输和管理。

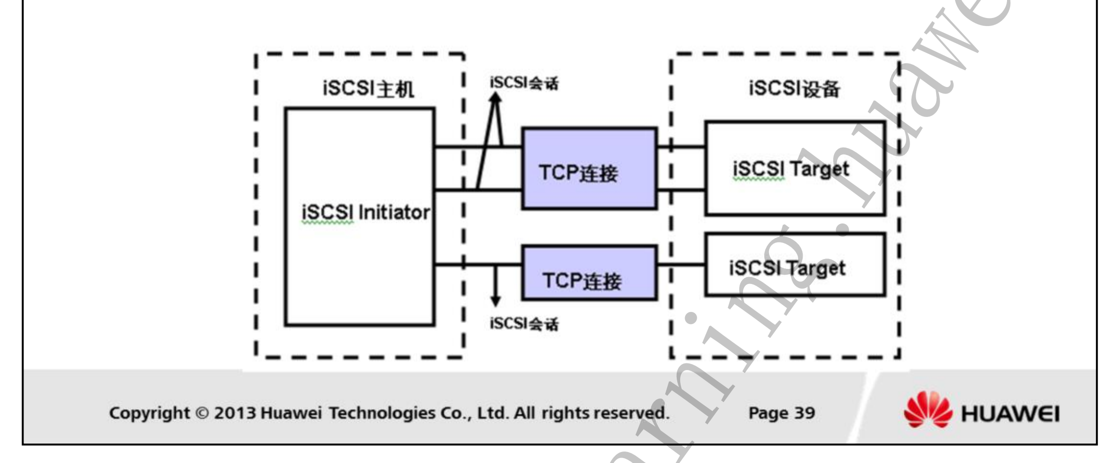


iSCSI协议位于TCP/IP协议和SCSI协议之间，可以起到连接这两种协议网络的作用。在物理层，iSCSI实现了对千兆以太网接口的支持，这使得所有支持iSCSI接口的系统都可以方便的直接连接到千兆以太网的路由器或者交换机上。iSCSI位于物理层和数据链路层之上，直接面向操作系统的标准SCSI命令集。

在iSCSI通信中，具有一个发起I/O请求的启动设备（Initiator）和响应请求并执行实际I/O操作的目标设备（Target）。在启动设备和目标设备建立连接后，目标设备在操作中作为主设备控制整个工作过程。在一般情况下将主机总线适配器（HBA）作为启动设备，磁盘/磁带作为目标设备。iSCSI使用iSCSI Name来唯一鉴别启动设备和目标设备。地址会随着启动设备和目标设备的移动而改变，但是名字始终是不变的。建立连接时，启动设备发出一个请求，目标设备接收到请求后，确认启动设备发起的请求中所携带的iSCSI Name是否与目标设备绑定的iSCSI Name一致，如果一致，便建立通信连接。每个iSCSI节点只允许有一个iSCSI Name，一个iSCSI Name可以被用来建立一个启动设备到多个目标设备的连接，多个iSCSI Name可以被用来建立一个目标设备到多个启动设备的连接。



在iSCSI启动器上用户发起了一个SCSI请求，操作系统将请求处理为一条或多条SCSI指令，由CPU或者是HBA卡对指令或数据进行封装形成一个iSCSI报文，然后传送给TCP/IP层，由TCP/IP协议把iSCSI报文封装成IP包并在网络中传输。当该报文到达目的端以后TCP/IP协议将数据包进行解封装，还原成一个iSCSI封装报文，再将iSCSI包还原为SCSI指令，交由操作系统处理。



iSCSI协议的会话就是在一个网络上封包和解包的过程。在网络的一端，数据包被封装成包括TCP/IP头、iSCSI识别包和SCSI数据三部分内容。当数据包被传输至网络另一端时，这三部分内容分别被有序的解封装，还原为原始的SCSI数据。iSCSI会话建立前必须先建立TCP连接，当TCP经过三次握手建立起连接之后才能建立iSCSI会话。一个TCP会话中可以包含一个或者多个iSCSI会话。

启动器设备可以通过下列方法发现目标设备。

- 在启动设备上设置目标设备的地址。
- 在启动设备上设置默认目标设备地址，启动设备可通过“Send Targets”命令从默认目标设备上获取iSCSI名字列表。
- 发出服务定位协议（Service Location Protocol，SLP）广播请求，等待目标设备回应。
- 查询存储设备名字服务器，获取可访问的目标设备列表。

iSCSI Discovery会话

- Discovery会话仅用于iSCSI Target discovery而建立的会话。

Initiator | Target

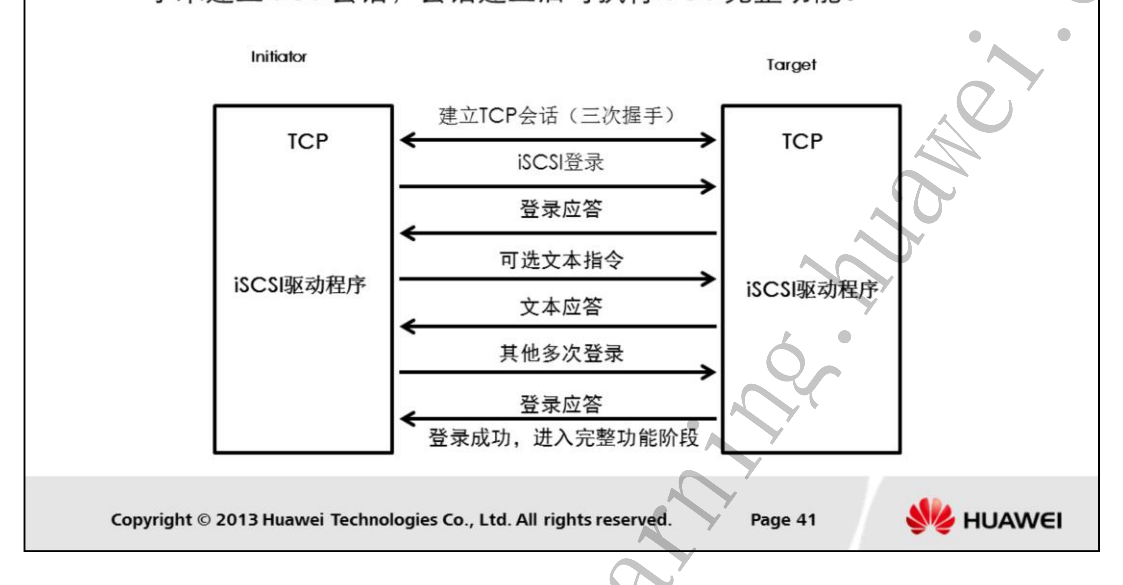
Initiator		Target
TCP	建立TCP会话（三次握手）	TCP
Login Request	Discovery Session	Login Response
Login Request	Parameter Session	Login Response
Text Request	Send Target	Text Response

 Page 40 

iSCSI有两种会话，分别是Discovery会话和Normal会话。Discovery会话仅用于iSCSI Target discovery而建立的会话。Normal会话是无限制会话，iSCSI无需执行Send Target命令发现请求，iSCSI Initiator直接使用iSCSI Target的名字来建立iSCSI会话，会话建立后可执行iSCSI完整功能。

- iSCSI的Discovery会话

在建立iSCSI会话前需要先建立TCP连接，TCP连接通过三次握手过程来建立。而Discovery会话的建立分为三个阶段，首先是Initiator和Target之间的登录参数协商阶段， Initiator发送Login Request报文请求登录，Target在收到请求信息后返回Login Response报文给Initiator，同意Initiator登录，从而完成初步的登录协商。在登录之后，传送数据之前还需要进行一次从参数的协商，这个过程被称为完整功能状态下的参数协商。最后再由Initiator发送Sent Target命令请求报文Text Request，Target端收到请求报文以后，查询到网络中存在的iSCSI信息后发送Text Response报文给Initiator，并返回一系列和它相连的iSCSI Target的信息，最终建立会话。



iSCSI normal会话分以下三个阶段:

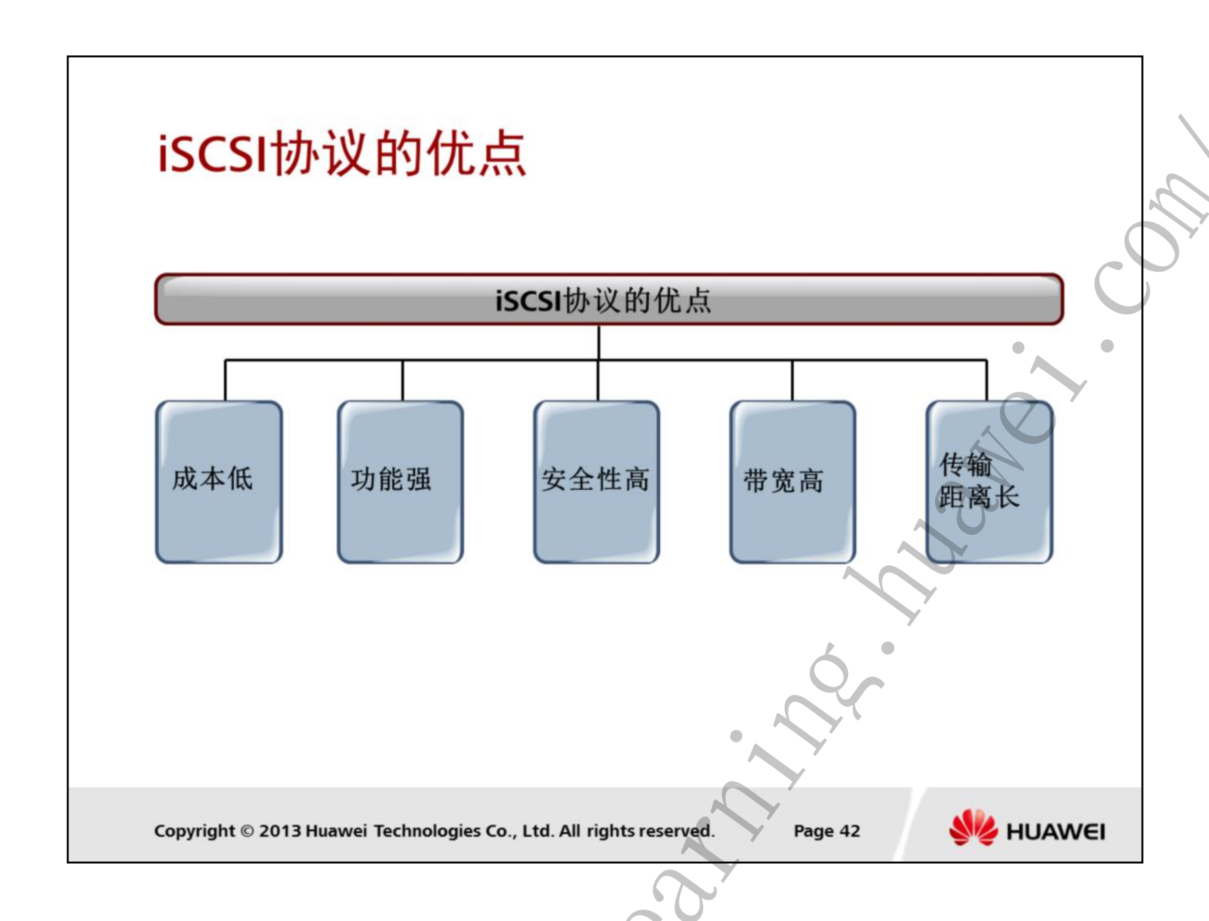
- 登录阶段：初始化登录阶段、安全认证阶段和操作协商
- 完整功能阶段
- 登出阶段

iSCSI的登录阶段等同于FC端口登录过程。该过程用来在两个网络实体间调整各个参数并确认登录的访问权限。如果iSCSI登录阶段成功完成，目标设备将确认启动设备的登录，否则登录将不确认，同时TCP连接中断。

登录一旦确认，iSCSI会话将进入完整功能阶段。如果建立了多个TCP连接，iSCSI将要求每个命令/响应对应一个TCP连接。但是，不同的数据传输可以在一个会话中通过不同的TCP连接。在数据传送端，启动器发送/接收最新的数据，而目标器在完成数据传输后发送确认响应。iSCSI注销命令用来结束一个会话，在出现连接错误的时候也会发送它，以实现连接中断处理。iSCSI登录是用来在启动设备和目标设备之间建立TCP连接的机制。登录的作用包括鉴别通信双方，协商会话参数，打开相关安全协议并且给属于该会话的连接作标志。

登录过程完成后，iSCSI会话进入全功能状态（Full Feature Phase），这时启动设备就能通过iSCSI协议访问目标设备里的各逻辑单元了。

iSCSI会话拆除时，Initiator首先向Target发送Logout Request请求报文，Target接受到请求报文后返回相应的Logout Response报文，至此，iSCSI会话可以拆除。在拆除iSCSI会话后还需要拆除TCP连接，TCP连接的拆除是通过四次握手来完成的。

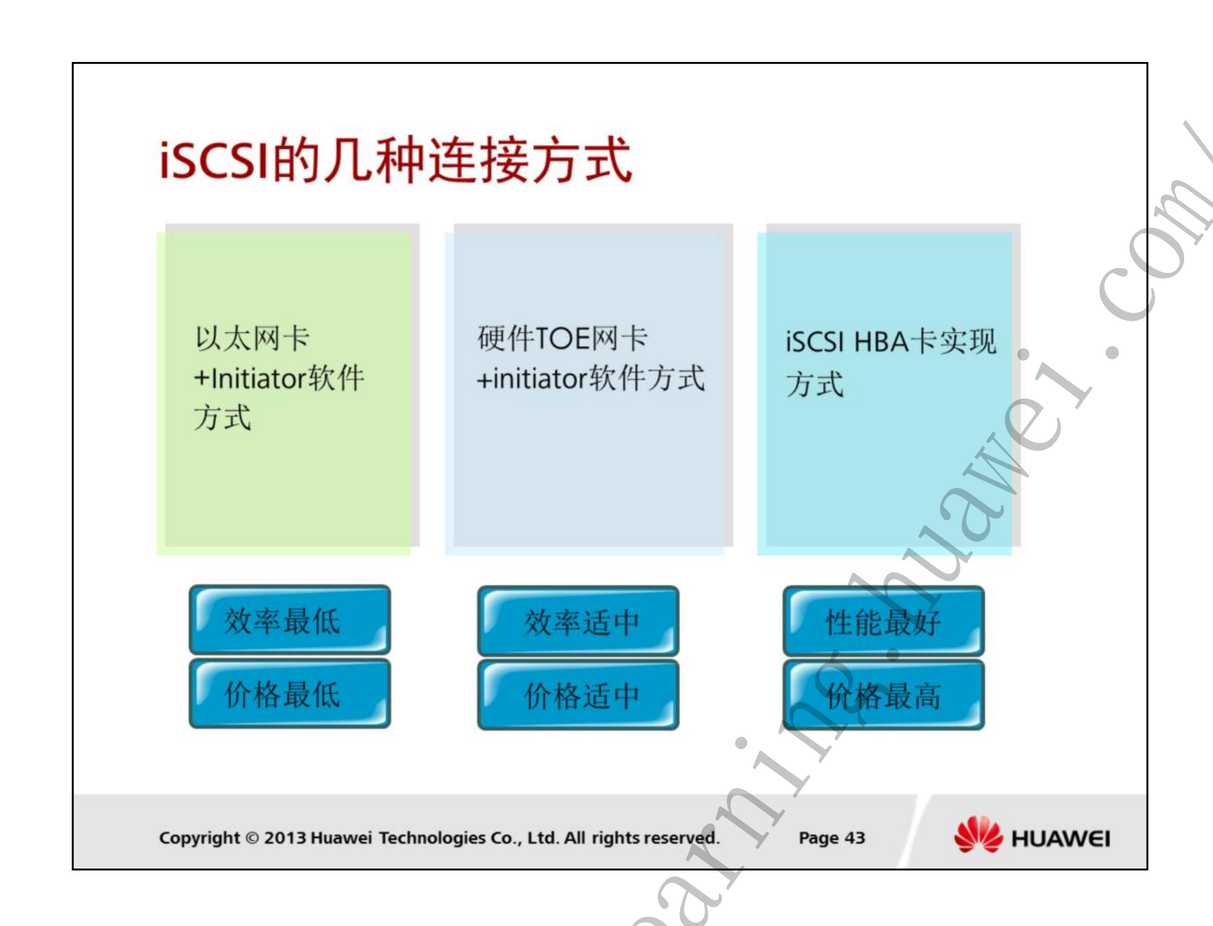


iSCSI协议与FC及其他协议相比具有一定的优势，也正因为如此，得到用户广泛的认可。与光纤通道相比，在连接距离上比FC-SAN强，它可突破FC-SAN目前10公里的极限，扩展到整个WAN上。另外，iSCSI更加经济，其成本的节约又体现在以下几个方面：

- 因为使用的是传统的IP，用户有良好的使用基础，所以在培训方面的费用可大大降低，而且也不必设立单独的岗位。
- iSCSI可利用现有的、容易理解的TCP/IP基础设施来构建SAN，网络部署成本也将大大降低。
- 随着千兆以太网的应用，用户将可得到传输速率为1Gbps的存储网络，而不需改变现有的基础设施，在维护和管理方面同样可降低成本。

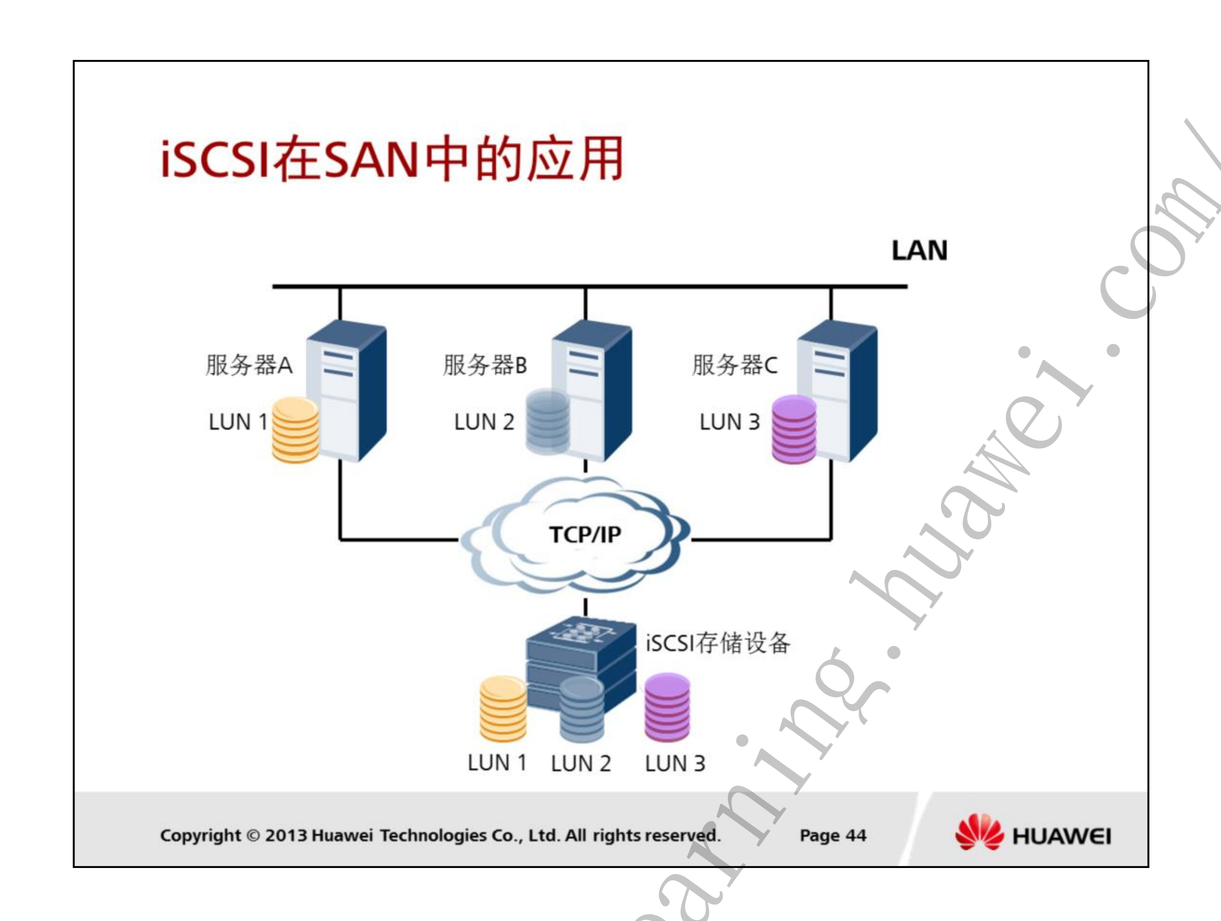
相对其他协议来说，iSCSI技术具有如下优势：

- 带宽高：随着技术的进步，IP网络的带宽的发展相当迅速，1Gbps的以太网早已大量占据市场，10Gbps以太网的应用也已开始启动。而且，该协议得到IBM、Cisco、Intel、Brocade和Adaptec等业界厂商的支持，发展前景良好。
- 可用性强：在技术实施方面，iSCSI以稳健、有效的IP及以太网架构为骨干，使网络的可用性大大增强。
- 功能强：完全解决了数据远程复制(Data Replication)及灾难恢复(Disaster Recovery)的难题。
- 安全性高：以往的FC-SAN及DAS大都是在封闭的环境内，安全要求相对较低。iSCSI却将这种概念颠倒过来，让存储的数据在互联网内流通，令用户感到需要提升安全要求，而iSCSI已内建了支持IP Sec的机制，并且在芯片层面执行有关指令，确保了安全性。



iSCSI设备的主机接口一般默认都是IP接口，可以直接与以太网络交换机和iSCSI交换机连接，形成一个存储区域网络。根据主机端HBA卡、网络交换机的不同，iSCSI设备与主机之间有三种连接方式。

- 以太网卡＋Initiator软件方式：采用通用以太网卡实现网络连接，主机CPU通过运行软件完成iSCSI层和TCP/IP协议栈的功能。由于采用标准网卡，因此这种方式的硬件成本最低。但主机的运行开销大大增加，造成主机系统性能下降。实验证明，当通信量增大时，主机CPU的利用率可达90%以上。
- 硬件TOE网卡＋Initiator软件方式：采用特定的智能网卡，iSCSI层的功能由主机来完成，而TCP/IP协议栈的功能由网卡来完成。与纯软件方式相比，部分降低了主机的运行开销。
- iSCSI HBA卡实现方式：iSCSI层和TCP/IP协议栈的功能均由主机总线适配器来完成，对主机CPU的需求最少。



使用iSCSI构建的IP-SAN和FC-SAN一样具备良好的扩展性和灵活性，可通过网络交换设备与多台主机连接。通过网络交换设备连接时，iSCSI存储上的LUN对于主机而言相当于裸设备，因此需要注意文件系统的管理问题。iSCSI设备上创建多个LUN，不同的LUN划分给不同的主机，使得各主机可以分别管理和访问自己的LUN这就相当于将网络中多个主机的本地磁盘集中放置在一个网络化的存储设备中，主机之间实现存储硬件设备的共享。



目录



什么是NAS？

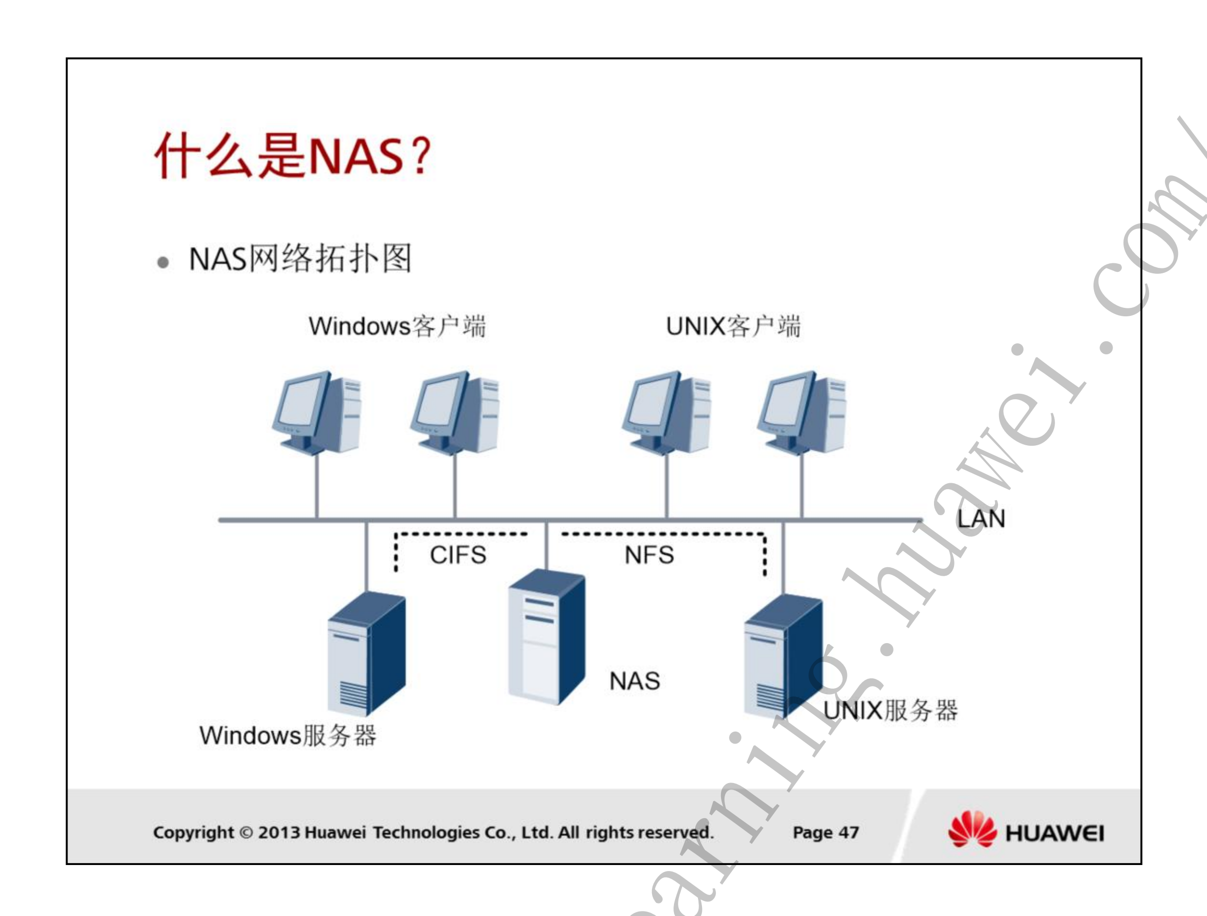
- 拥有可访问的磁盘阵列
- 拥有文件系统
- 对外提供访问文件系统的接口





NAS是Network Attached Storage的简称，中文称为网络附加存储。按字面简单说是连接在网络上，具备资料存储功能的设备，因此也称为“网络存储器”或者“网络磁盘阵列”。在NAS存储结构中，存储系统不再通过I/O总线附属于某个服务器或客户机，而直接通过网络接口与网络直接相连，由用户通过网络访问。 NAS实际上是一个带有瘦服务器的存储设备，其作用类似于一个专用的文件服务器。这种专用存储服务器去掉了通用服务器原有的不适用的大多数计算功能，而仅仅提供文件系统功能。与传统以服务器为中心的存储系统相比，数据不再通过服务器内存转发，直接在客户机和存储设备间传送，服务器仅起控制管理的作用。

一套简单的NAS系统可以由一台标准的运行Linux或者微软WSS（Windows Server）的服务器构成。



NAS本身能够支持多种协议（如NFS、CIFS、FTP、HTTP等），而且能够支持各种操作系统。通过任何一台工作站 ， 采用IE或Netscape浏览器就可以对NAS设备进行直观方便的管理。

NAS和SAN最大的区别就在于NAS有文件操作和管理系统，而SAN却没有这样的系统功能，其功能仅仅停留在文件管理的下一层，即数据管理。SAN和NAS并不是相互冲突的，是可以共存于一个系统网络中的，但NAS通过一个公共的接口实现空间的管理和资源共享，SAN仅仅是为服务器存储数据提供一个专门的快速后方通道。

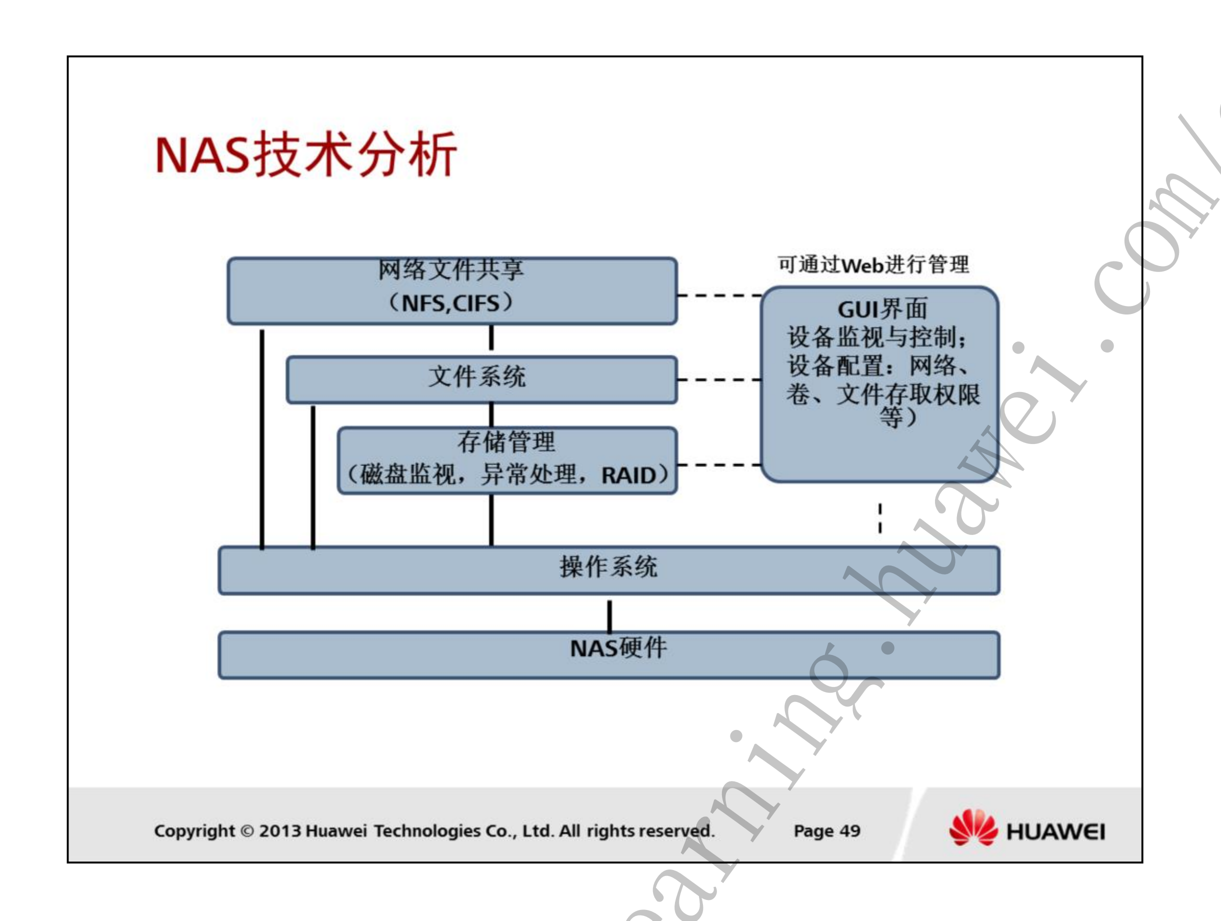
为什么FTP文件服务不属于NAS？FTP只能将文件传输到本地的目录之后才能执行，而网络文件系统可以允许直接问源端的文件，不需要将数据复制到本地再访问。

NAS技术分析

存储管理模块	OS文件系统模块	网络模块
• RAID • FC • SAS	• 操作系统 • 文件系统	• NFS • CIFS

Page 48
HUAWEI

- NAS的主要技术包括三个方面：存储管理模块、OS文件系统模块、网络模块。
    - 存储管理模块是指后端存储部分，这部分功能模块提供了真正的物理存储空间，主要技术是RAID、SCSI、SAS、FC。
    - OS文件系统是指NAS引擎部分，这部分提供了NAS底层所是使用的文件系统，以及承载文件系统、各种前端协议的操作系统。
    - 网络模块提供了和用户交互的网络协议，主要包括NFS和CIFS，用户最终通过这些协议访问存储空间。



NAS系统软件设计的基本要求是较高的稳定性和I/O吞吐率，并能满足数据共享、数据备份、安全配置、设备管理等要求。该结构分为五个模块：操作系统、存储管理器、文件系统、网络文件共享和GUI管理模块。

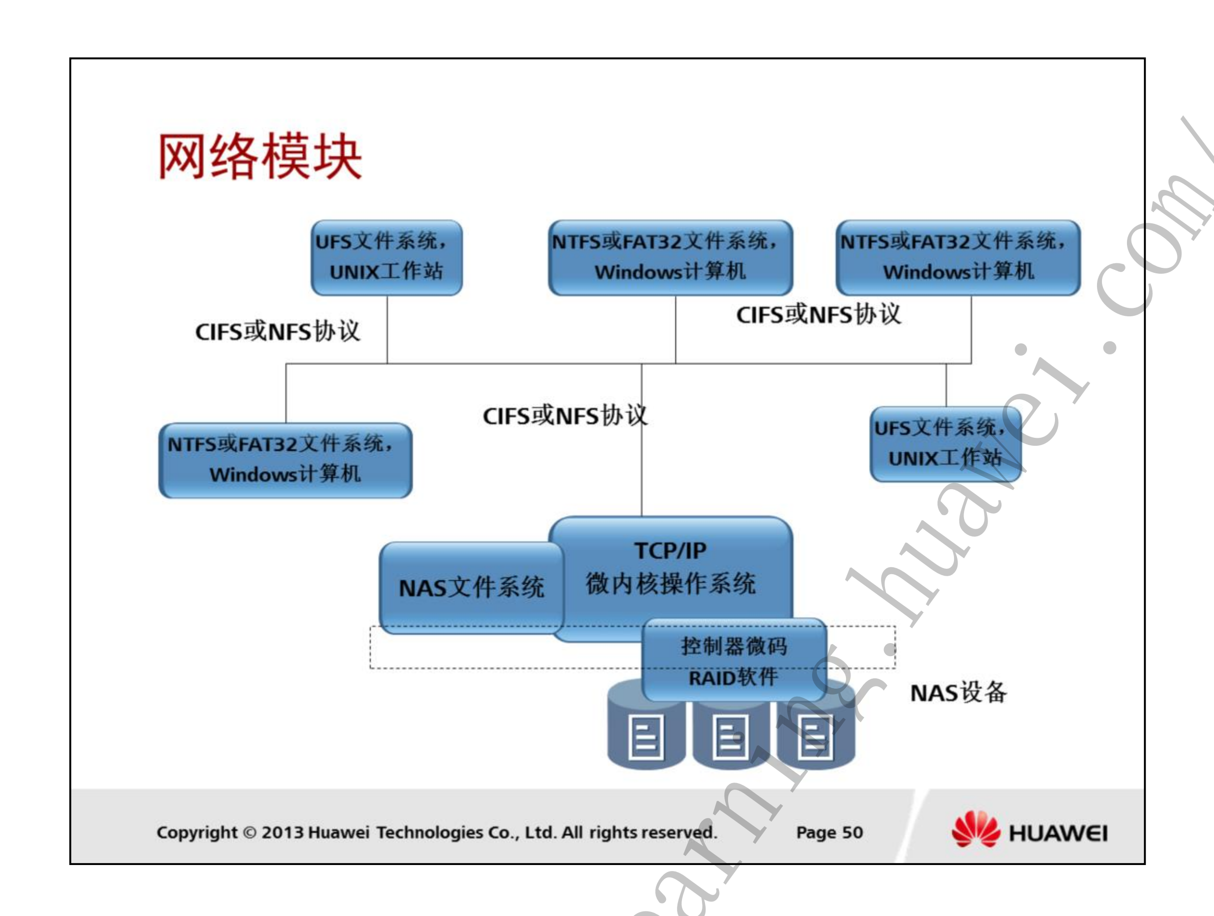
鉴于Linux、FreeBSD等免费的开放源码操作系统具有稳定、可靠、高效的优秀特性，现在大部分NAS设备都是基于此类操作系统开发的。

存储管理器的主要功能是磁盘和分区的管理，主要包括磁盘的监测与异常处理和逻辑卷的配置管理，一般应支持磁盘的热插拔、热替换等功能和RAID0、RAID1、RAID5类型的逻辑卷。存储管理器实现简化的、集中的存储管理功能，保证数据的完整性，并增强数据的可用性。

文件系统提供持久性存储和管理数据的手段，它必须是32位或以上并能支持多用户，应具备日志文件系统功能以使系统在崩溃或掉电重启后能迅速恢复文件系统的一般性和完整性，进一步提供NAS的可用性。此外，文件系统还应具有快照功能。快照不仅能恢复被用户错误修改或删除的文件，而且能实现备份窗口为零的文件系统活备份。

网络文件共享一般支持以下一些文件传输和共享协议，如FTP和HTTP协议、Unix系统的NFS、Windows系统的CIFS、Novell系统的NCP（Novell core protocol）、Apple系统的AFP（appletalk file protocol）等，因此NAS设备具有较好的协议独立性。

GUI管理提供给系统管理员一个友好的界面，使之仅通过web浏览器操作就能远程监视和管理NAS设备的系统参数，如：网络配置、用户与组管理、卷以及文件共享权限等。



NAS设备中所包含的标准文件系统可以对公用互联网文件系统（CIFS）或是网络文件系统（NFS）提供支持，也有可能同时支持二者。在许多情况下，它都使用标准的网络文件系统来作为NAS专用文件系统的接口。大多数NAS设备需要用这种方式来管理其自身的存储资源。

NFS（网络文件系统）是Unix系统间实现磁盘文件共享的一种方法，支持应用程序在客户端通过网络存取位于服务器磁盘中数据的一种文件系统协议。其实它包括许多种协议，最简单的网络文件系统是网络逻辑磁盘，即客户端的文件系统通过网络操作位于远端的逻辑磁盘。现一般在Unix主机之间采用Sun开发的NFS（Sun），它能够在所有Unix系统之间实现文件数据的互访，逐渐成为主机间共享资源的一个标准。

CIFS是由微软开发的，用于连接Windows客户机和服务器。经过Unix服务器厂商的重新开发后，它可以用于连接Windows客户机和Unix服务器，执行文件共享和打印等任务。它最早的由来是NetBIOS，这是微软开发的在局域网内实现基于Windows名称资源共享的API。之后，产生了基于NetBIOS的NetBEUI协议和NBT(NetBIOS OVER TCP/IP)协议。NBT协议进一步发展为SMB（Server Message Block Potocol）和CIFS（Common Internet File System，通用互联网文件系统）协议。其中，CIFS用于Windows系统，而SMB广泛用于Unix和Linux，两者可以互通。SMB协议还被称作LanManager协议。CIFS支持与SMB的服务器通信而实现共享。微软操作系统家族和几乎所有Unix服务器都支持SMB协议/SAMBA软件包。

CIFS和NFS的对比：CIFS面向网络连接的共享协议，对网络传输的可靠性要求高，常使用TCP；NFS是独 立于传输的，可使用TCP或UDP。



目录

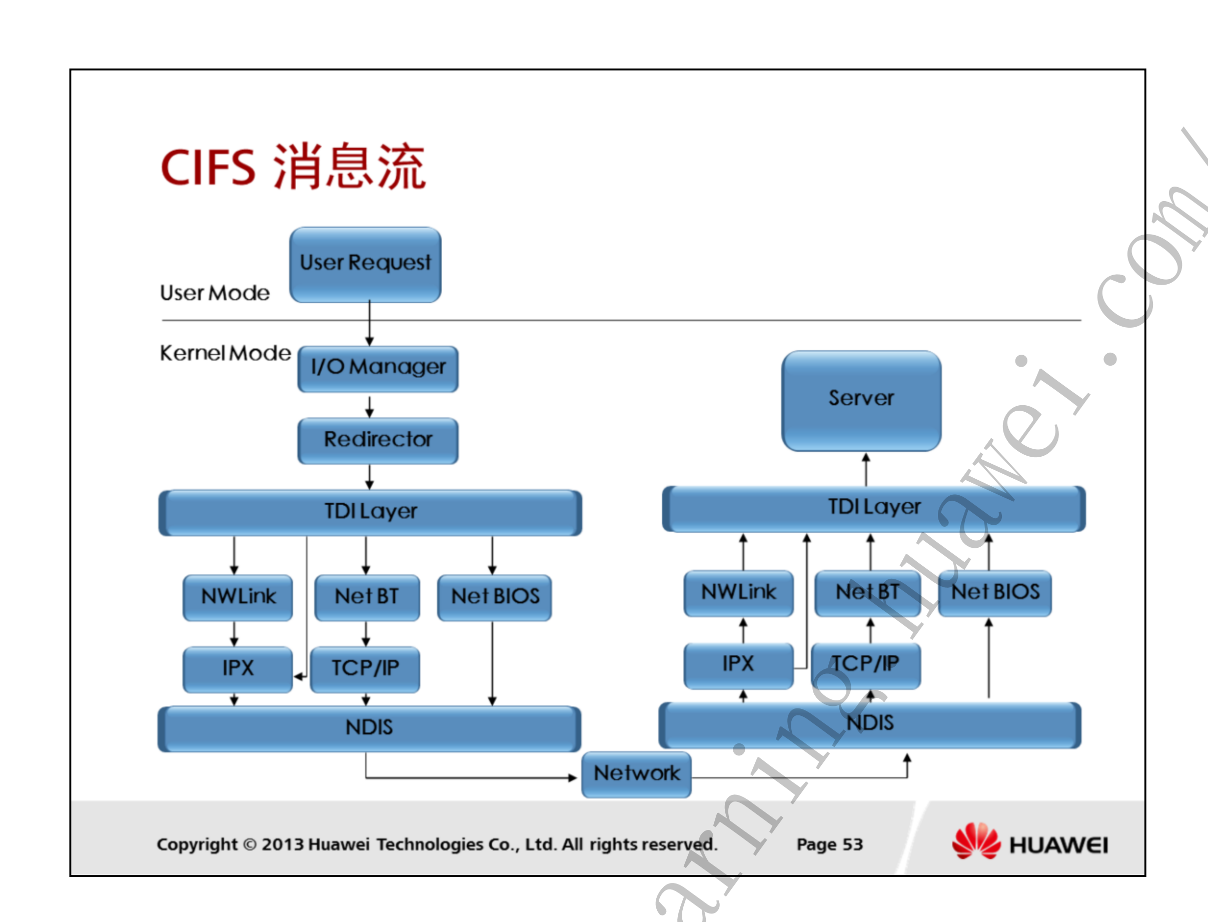


CIFS是什么

- **CIFS (Common Internet File System)**，通用Internet文件系统 ，一个新提出的协议，它使程序可以访问远程Internet计算机上的文件并要求此计算机的服务。
- CIFS是公共的或开放的SMB协议版本，包含主要模块有NBT、SMB、Browsing

 Page 52 

CIFS使用客户/服务器模式，客户程序请求远在服务器上的服务器程序为它提供服务，服务器获得请求并返回响应。


CIFS 消息流
User Request
User Mode
Kernel Mode
I/O Manager
Redirector
TDI Layer
NWLink
Net BT
Net BIOS
IPX
TCP/IP
NDIS
Server
TDI Layer
NWLink
Net BT
Net BIOS
IPX
TCP/IP
NDIS
Network
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 53
HUAWEI



目录



NFS主要模块：RPC、鉴权、传输

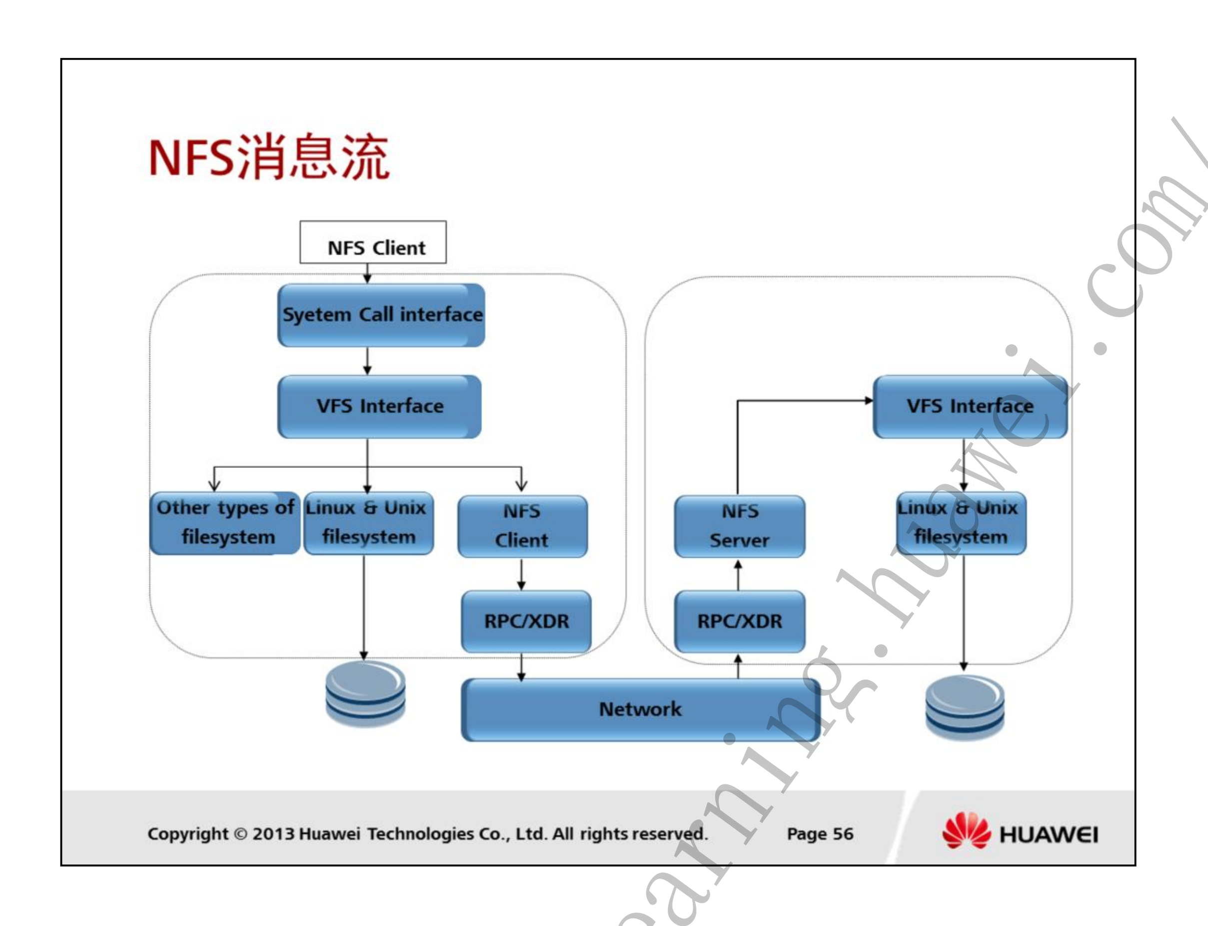
- NFS（Network File System）–网络文件系统
  - NFS（网络文件系统）是Unix系统间实现磁盘文件共享的一种方法，支持应用程序在客户端通过网络存取位于服务器磁盘中数据的一种文件系统协议。

 Page 55



NFS (Network File System，网络文件系统)是当前主流异构平台共享文件系统之一。主要应用在UNIX环境下。 最早是由SUN microsystem开发，现在能够支持在不同类型的系统之间通过网络进行文件共享，广泛应用在FreeBSD、SCO、Solaris等等异构操作系统平台， 允许一个系统在网络上与它人共享目录和文件。通过使用NFS，用户和程序可以象访问本地文件一样访问远端系统上的文件，使得每个计算机的节点能够像使用本地资源一样方便地使用网上资源。换言之，NFS 可用于不同类型计算机、操作系统、网络架构和传输协议运行环境中的网络文件远程访问和共享。

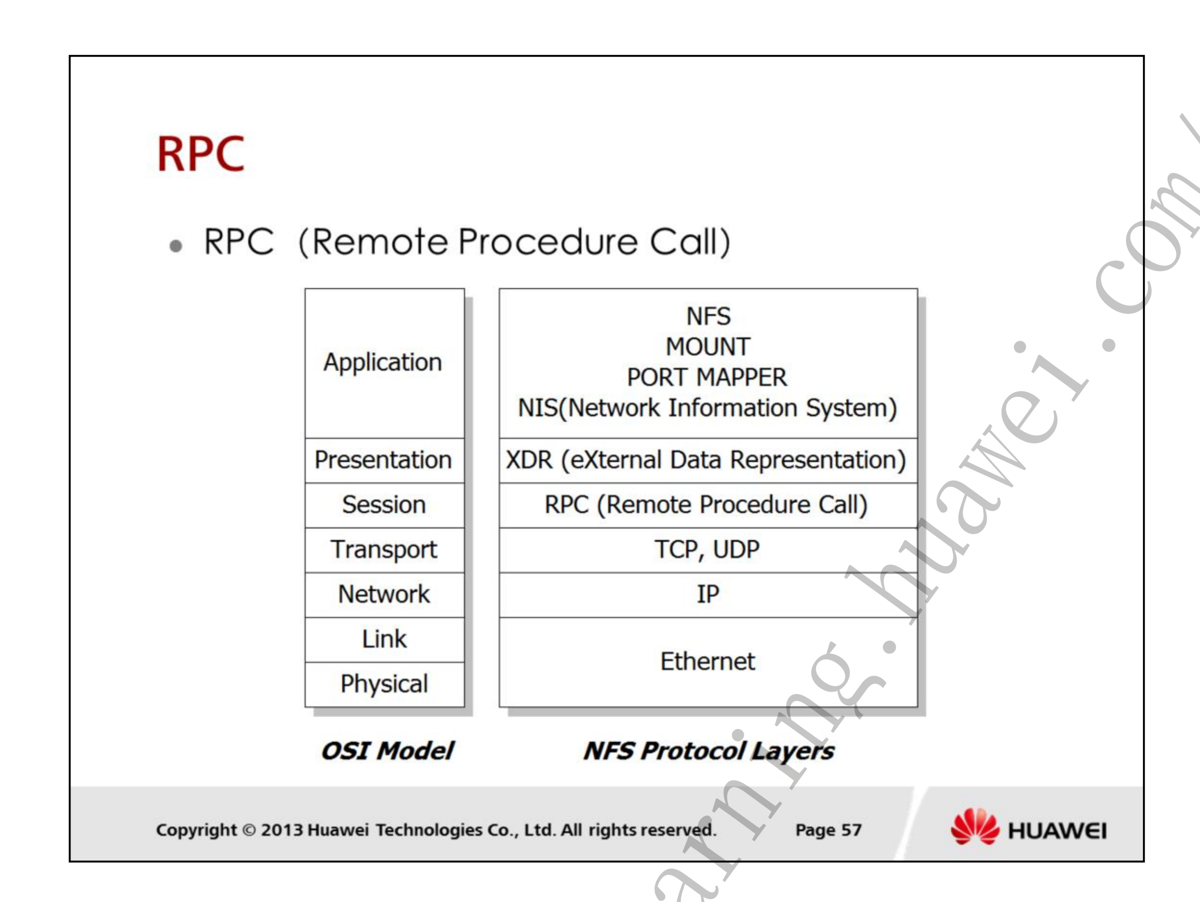
NFS的工作原理是使用客户端/服务器架构，由一个客户端程序和服务器程序组成。服务器程序向其它计算机提供对文件系统的访问，其过程就叫做“输出”。NFS 客户端程序对共享文件系统进行访问时，把它们从 NFS 服务器中“输送”出来。文件通常以“块”为单位进行传输. 其尺寸是 8K (虽然它可能会将操作分成更小尺寸的分片).NFS 传输协议用于服务器和客户机之间文件访问和共享的通信，从而使客户机远程地访问保存在存储设备上的数据。



NFS工作流程，当NFS客户端发送请求访问远端NFS服务器上的文件时，内核会通过VFS层，将请求通过RPC发送到远端的NFS服务器，RPC可以通过TCP/UDP来进行传输，然后NFS服务器在2049端口接收到NFS客户端的RPC请求，并将其转发到NFS服务器的本地文件系统。

NFS服务器在请求完成后，会将结果通过RPC发送回执到NFS客户端，其中有两点需要注意：

- NFS服务器发送回执到NFS客户端需要一定的时间，此时NFS服务器同时仍然能响应其他的NFS客户端请求，此时服务器上存在多个NFSD的进程。
- 在NFS客户端也存在相同的情况，NFS客户端会存在多个biod进程，用于发送NFS请求。



- RPC（Remote Procedure Call）
  - NFS本身不提供信息传输的协议和功能，而是使用了一些其它的传输协议。而这些传输协议用到这个RPC功能的。可以说NFS本身就是使用RPC的一个程序。或者说NFS也是一个RPC SERVER.所以只要用到NFS的地方都要启动RPC服务。这样SERVER和CLIENT才能通过RPC来实现PROGRAM PORT的对应。可以这么理解RPC和NFS的关系：NFS是一个文件系统，而RPC是负责负责信息的传输。
  - 网络七层协议与NFS协议的对应关系：NFS是处于应用层的协议，NFS的相关模块为mount，portmapper模块，mount进程可以对客户端的请求进行认证，portmaaper提供了客户端要连接的端口，NIS提供了一种域的认证方式，NFS及其相关协议都需要使用到RPC，所有的NFS请求都需要通过RPC来完成，XDR是RPC的一种数据编码格式。RPC可以使用TCP/UDP进行传输。

传输方式

- 使用UDP协议
  - UDP传输在NFS局域网的应用中传输速度快
  - UDP协议传输的开销小
- 使用TCP协议
  - 可靠性高，有效的阻塞控制
  - 客户端和服务器端都保留TCP连接的状态
  - 服务器崩溃时，客户端只需要打开一个新的TCP连接
  - 客户端崩溃时，服务器端在新的TCP连接到来时，关闭原来的TCP连接



不论NFS使用UDP还是TCP，NFS连接都是无状态的。



思考题

1. SCSI协议和FC，SAS等协议的关系是什么？
2. SAN和NAS的区别是什么？





总结

- 存储系统的基本构架；
- 存储系统的发展趋势；
- 存储系统常用协议和发展趋势。





习题

- 判断题
  1. 使用iSCSI HBA的主机仍然需要使用Initiator软件。（T of F)
- 多选题
  1. 存储网络的硬件组件有哪些?（　　）

     A. HBA

     B. Switch

     C.存储设备

     D.主机

 Page 61 

- 习题答案：
  - 判断题：1.F
  - 多选题：1.ABC

附录

- SCSI： （Small Computer System Interface，小型计算机系统接口）是一种为小型机研制的接口技术，用于主机与外部设备之间的连接。SCSI-3是所有存储协议的基础，其它存储协议都用到SCSI的指令集。优点：与主机无关、多设备并行、高带宽。缺点：允许连接设备数量少、连接距离非常有限。
- SAS: （Serial Attached SCSI）即SCSI总线协议的串行标准，即串行连接SCSI；SAS采用串行技术以获得更高的扩充性，并兼容SATA盘。目前SAS的最高传输速率高达3Gpbs、6Gbps，支持全双工模式。
- FC: FC是光纤通道（Fiber Channel）的简称，用于服务器与共享存储设备的连接，存储控制器和驱动器之间的内部连接，是一种高性能的串行连接标准。其接口传输速率目前有4Gbps 、8Gbps几种标准。传输介质可以选择铜缆或光纤，传输距离远，支持多种互联拓扑结构。光纤通道是构建FC SAN的基础，是FC SAN系统的硬件接口和通信接口。
- iSCSI：（Internet Small Computer System Interface）互联网小型计算机系统接口，是一种在TCP/IP上进行数据块传输的标准，可以理解为SCSI over IP。iSCSI可构成基于IP的SAN，为用户提供高速、低价、长距离的存储解决方案。iSCSI将SCSI命令封装到TCP/IP数据包中，使I/O数据块可通过IP网络传输，是未来的发展之路。



Thank you

www.huawei.com

HC120920002 统一存储技术及应用




HC120920002
统一存储技术及应用
www.huawei.com
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



目标

学习完本课程后，您将能够：

- 熟悉RAID2.0技术
- 熟悉Smart技术
- 熟悉Hyper技术
- 链路管理和组网

 Page 1





目录

 Page 2

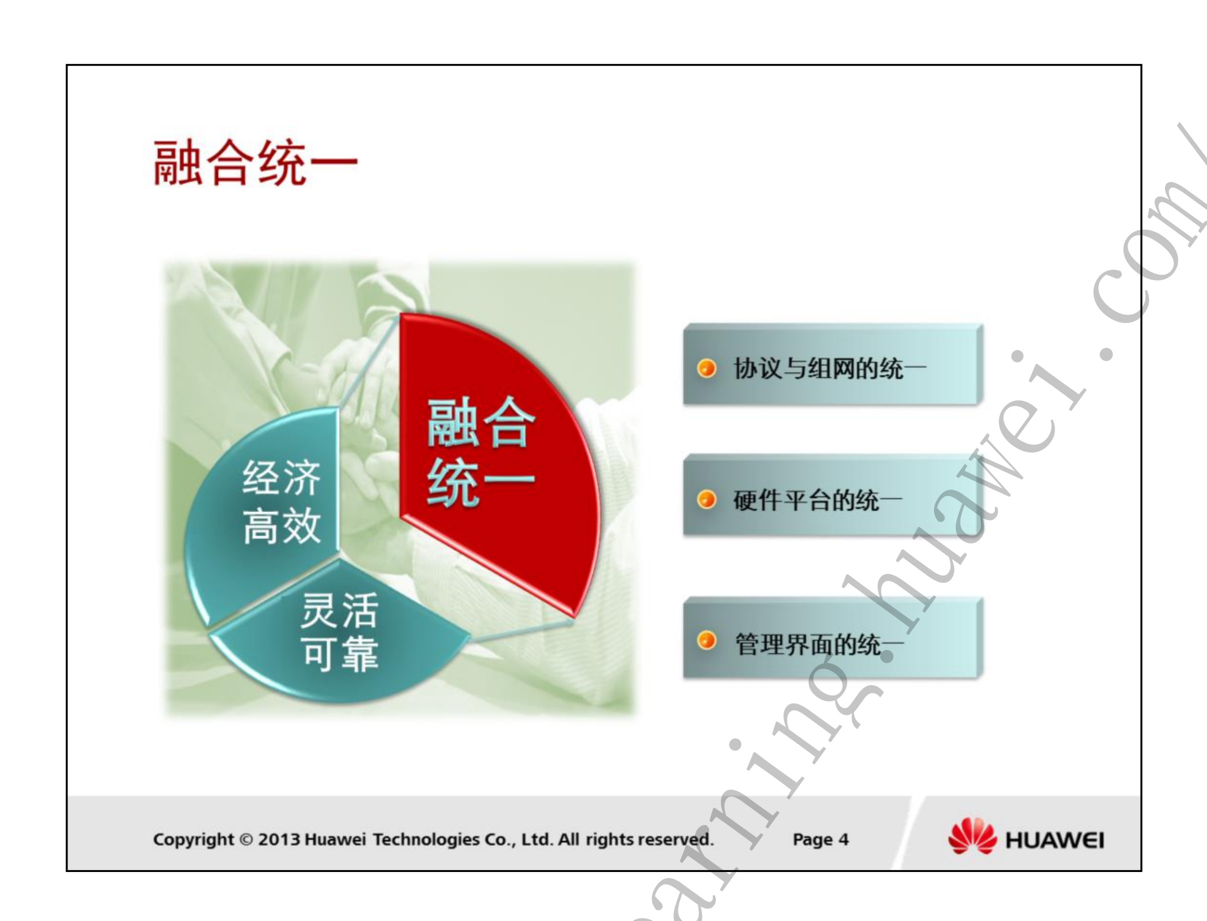


HUAWEI



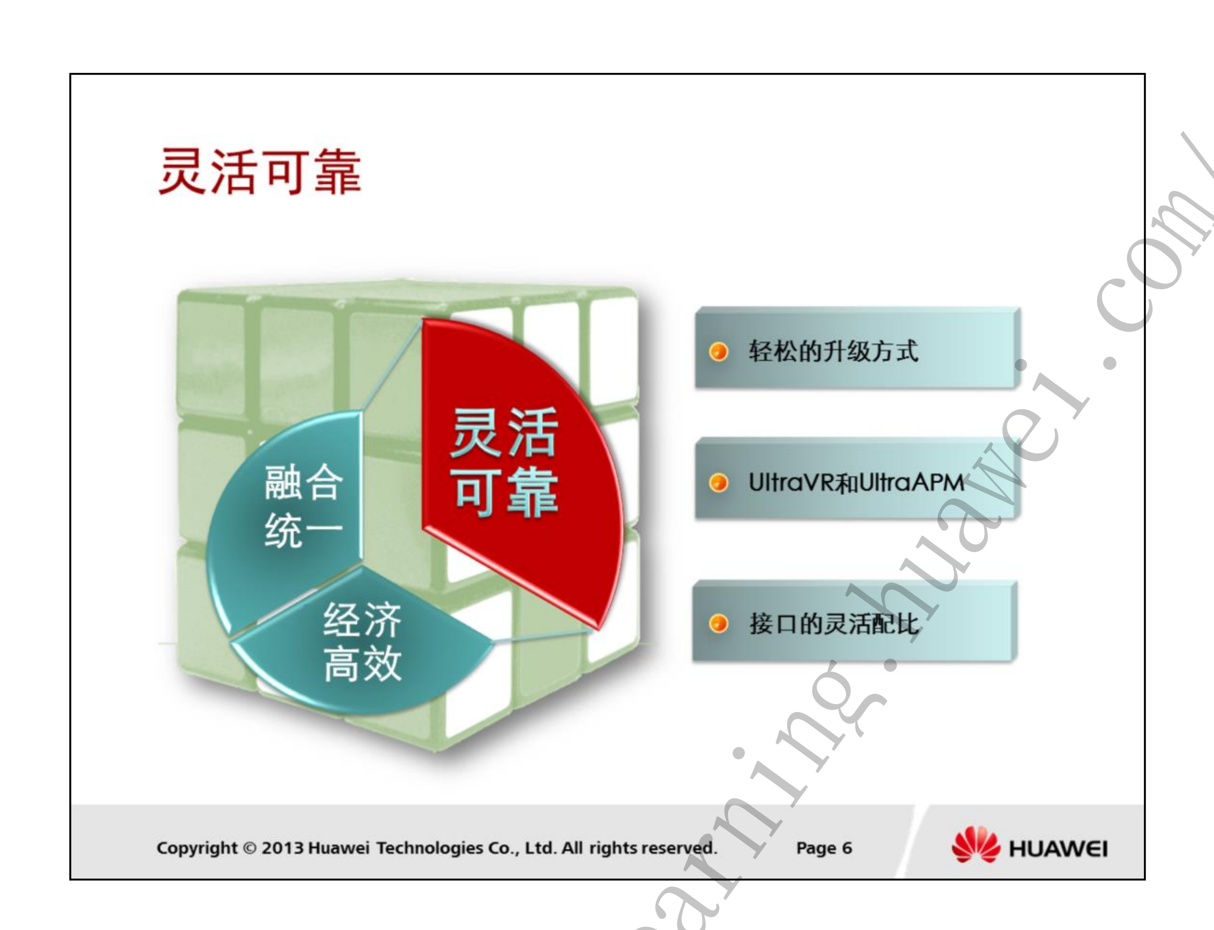
OceanStor 18000系列存储系统采用创新的智能矩阵架构 Smart matrix，通过Scale-out横向扩展，可为企业提供一至八个系统机柜和最多两个硬盘柜的存储系统。

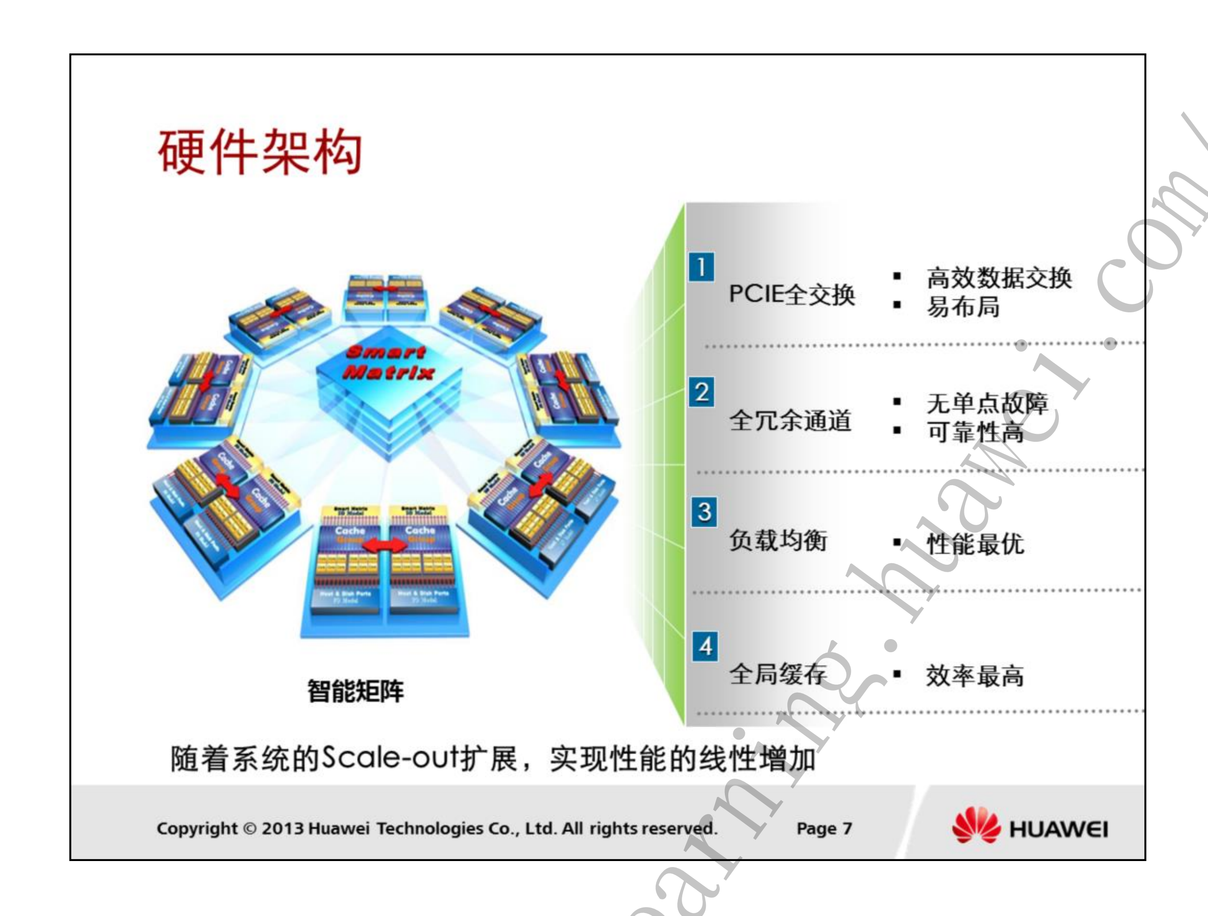
其中，OceanStor 18800最大能提供8个引擎，合16个控制器，3T缓存，192个主机端口，3216个硬盘的高扩展配置能力。OceanStor 18500与18800的区别在于硬件规格的区别。


融合统一
融合
统一
经济
高效
灵活
可靠
协议与组网的统一
硬件平台的统一
管理界面的统一
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 4
HUAWEI

经济高效

 Page 5 HUAWEI


灵活可靠
融合
统一
灵活
可靠
经济
高效
轻松的升级方式
UltraVR和UltraAPM
接口的灵活配比
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 6
HUAWEI



OceanStor 18000高端智能存储系统的Smart matrix架构采用了以下技术：

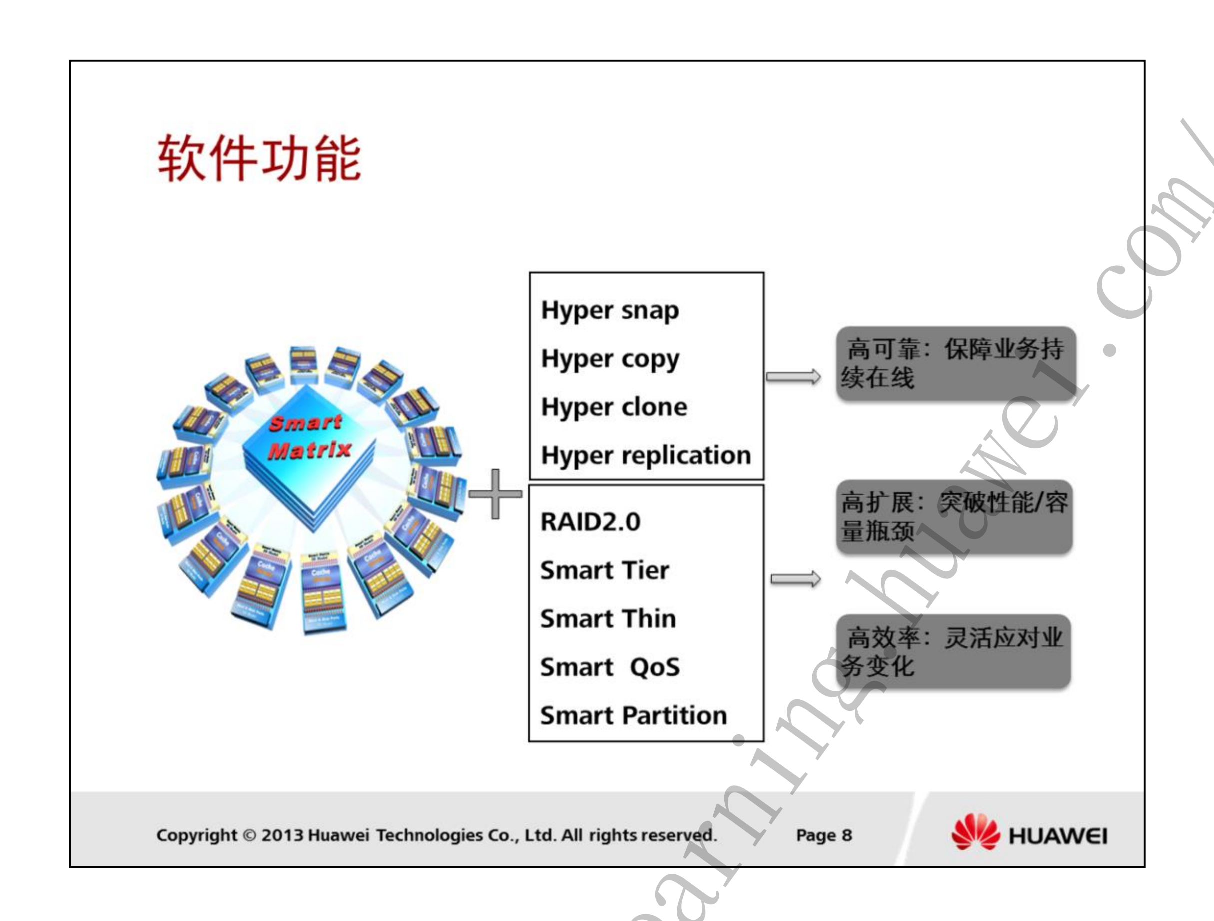
1、PCIE全交换技术：OceanStor 18000的每个引擎包含两个互为一对的控制器，这两个控制器间采用PCIe2.0*8背板镜像通道交换数据；所有控制器间采用PCIe2.0*4光缆连接到两个冗余的数据交换机中进行数据交换。这种PCIe全交换互联的设计，可以实现更加高效的数据交换，另外一方面对机柜之间的距离要求减少，有利于客户机房内部的机柜布局设计；

2、全冗余通道设计：不仅刚才讲到的PCIe数据交换通道，还有前端主机IO通道，后端硬盘通道，控制数据GE交换通道，OceanStor 18000都采用全冗余组网，无单点故障，可靠性高。

3、负载均衡技术：OceanStor 18000系统自动将不同LUN均衡分配到不同控制器，LUN空间均衡分散分布到系统内所有硬盘，从而使得不同控制器业务、硬盘压力相对均衡，配合华为自研多路径UltraPath选择最优路径下发IO，使系统性能达到了最优。

4、全局缓存技术：在控制器内，缓存分为读、写、镜像三种类型，可共用所有缓存；控制器间，各控制器可以访问其他所有控制器的缓存。缓存利用率最大，效率最高。

通过Smart Matrix的这些技术，使得OceanStor 18000高端智能存储，能随着系统的Scale-out扩展，实现性能的线性增加，甚至可达业界最高带宽192GB/s。



软件上，OceanStor 18000提供的Hyper snap、Hyper copy、Hyper clone、Hyper replication

多种高可靠的数据保护功能，保障了用户业务的持续在线；更提供了RAID2.0，smart tier，smart thin，smart QoS等高扩展高效率的特性，帮助用户灵活应对业务的变化，突破性能/容量瓶颈。



目录

1. **统一存储技术**



传统RAID

- RAID是一种将多个物理的硬盘组合成一个逻辑硬盘来使用的虚拟化技术。

RAID级别	描述
RAID0	数据条带化，无校验
RAID1	数据镜像，无校验
RAID5	数据条带化，校验信息分布式存放
RAID6	数据条带化，分布式校验提供两级冗余
RAID10	组内做RAID1，组间做RAID0，既采用数据镜像有做条带化



RAID 2.0原理

- 将硬盘划分成若干个连续的固定大小的存储空间，称为存储块，即chunk，或简称CK。
- Chunk按RAID策略组合成RAID组，称为存储块组，即chunk group，或简称CKG。
- 在CKG中划分若干小数据块，即extent。LUN就是由来自不同CKG的extent组成。
- 用作热备空间的CK也是分散在各个盘上的。

Page 11

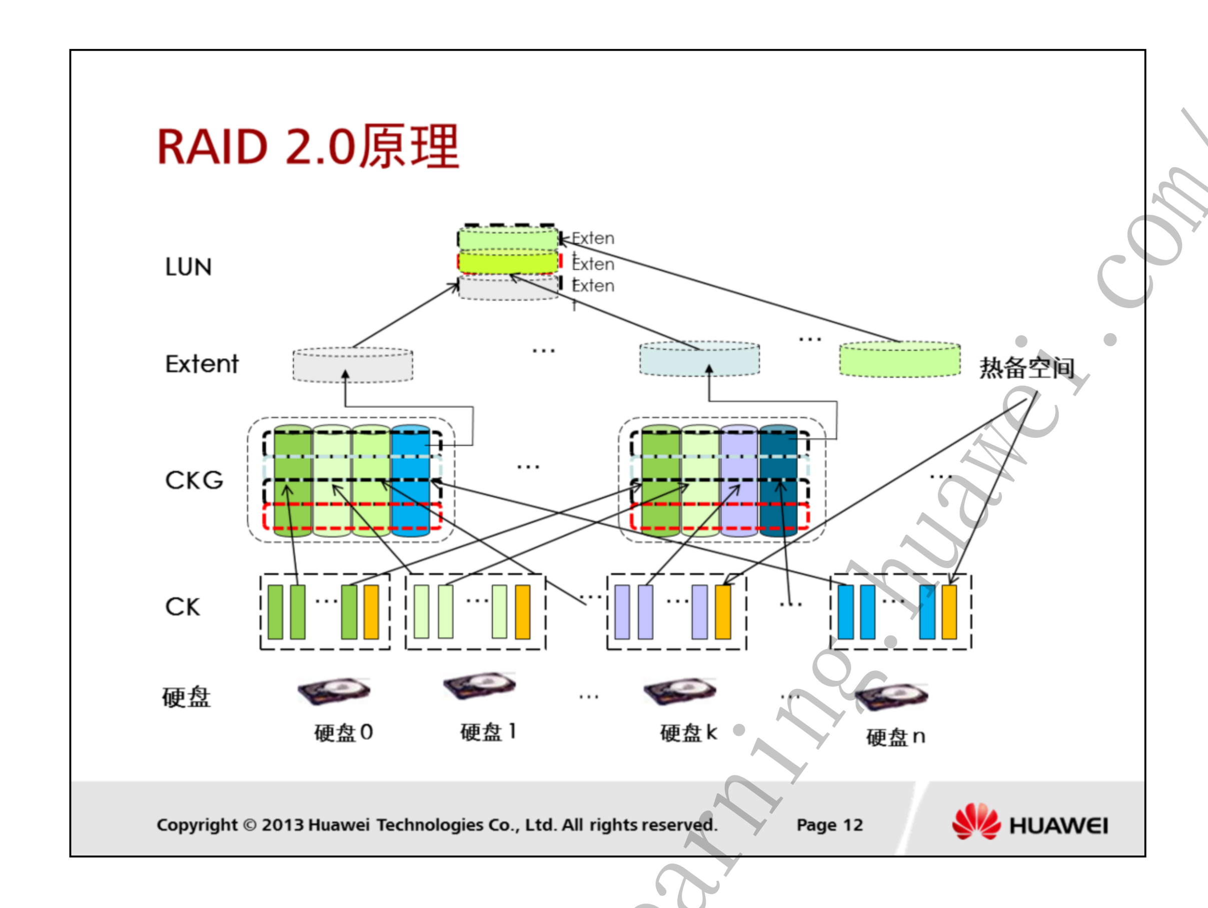


RAID2.0的核心思想是块级虚拟化。块级虚拟化的含义是：将系统中的硬盘划分成若干个连续的固定大小的存储空间，称为存储块，即chunk，或简称CK。图中带颜色的数据块即为CHUNK。

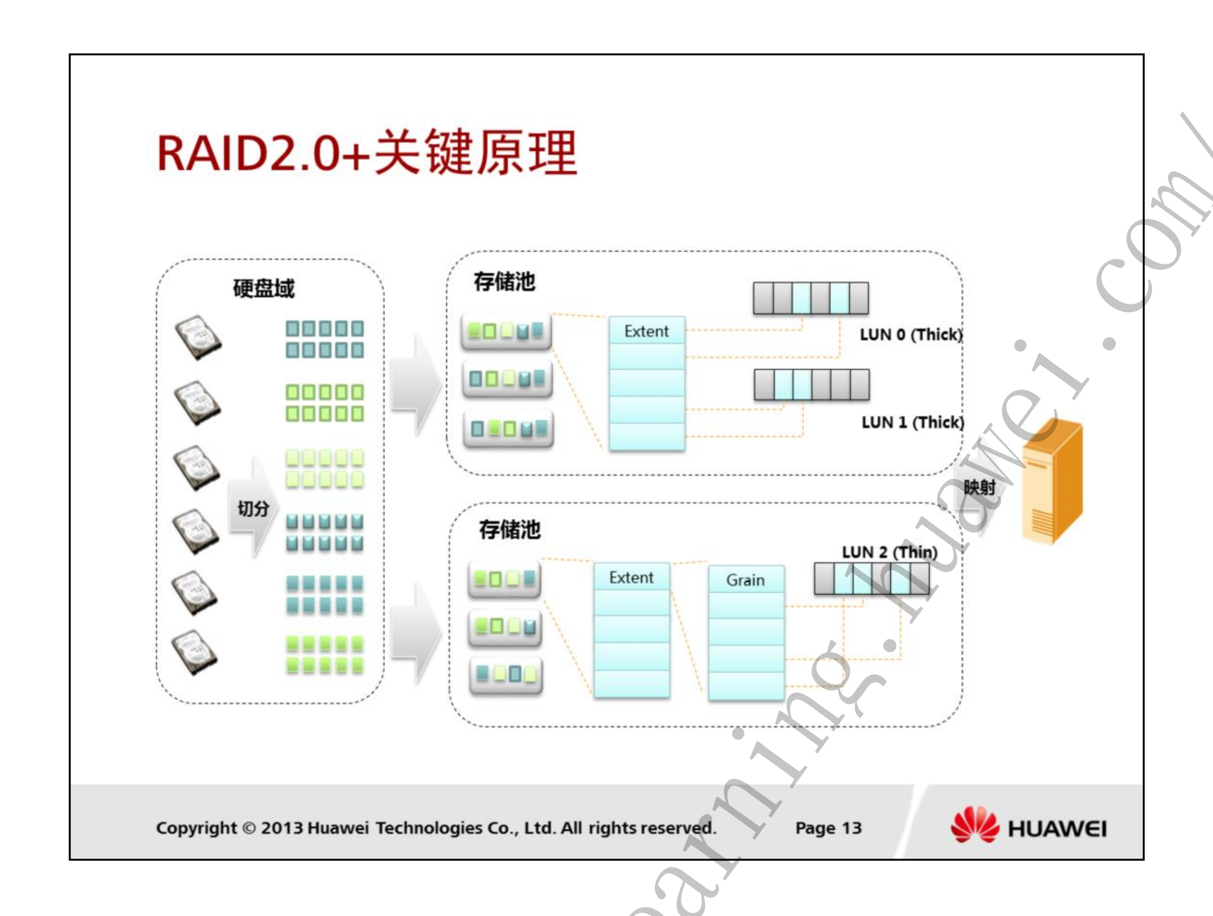
Chunck按RAID策略组合成RAID组，称为存储块组，即chunk group，或简称CKG。 在CKG中划分若干小数据块，即extent。LUN就是由来自不同CKG的extent组成。用作热备空间的CK也是分散在各个盘上的。

这里的extent数据块，也是OceanStor 18000存储系统分配LUN空间的最小单位。Extent来自不同的CKG，CKG中的CHUNK来自多个硬盘，可以是几十，上百个甚至上千个盘，而不是像传统RAID中来自故固定的几个盘，从而将LUN的空间充分分散分布到系统中更多的硬盘上。RAID2.0技术将物理空间和数据空间分散分布成分散的块，可以充分发挥系统的读写能力，方便扩展，也方便了空间的按需分配，数据的热度排布，迁移，它是所有Smart软件特性的实现基础。

同时，由于热备空间也是按CHUNK分散在多个盘上的，因此多个CKG的重构几乎可以同时进行，避免了写单个热备盘造成的性能瓶颈，大大减少了重构时间。



- Thick LUN以Extent为单位映射到LUN
- Grain在Extent的基础上进行更细粒度的划分
- Thin LUN以Grain 为单位映射到LUN



- 硬盘域内每个硬盘被切分为固定64MB的块（CK）
- 硬盘域内同种类型的硬盘被划分为一个个的Disk Group（DG），从同一个DG上随机选择多个硬盘，每个硬盘选取一个CK按照RAID算法组成Chunk Group（CKG）
- CKG被划分为固定大小的Extent
- Thick LUN以Extent为单位映射到LUN
- Grain在Extent的基础上进行更细粒度的划分
- Thin LUN以Grain 为单位映射到LUN

RAID2.0+软件逻辑对象

Disk Domain
Storage Pool
LUN
Volume
Grain Grain Grain
Extent Extent Extent
Zone Zone Zone
CKG CKG CKG
CK CK CK
LD LD LD
SSD SAS NL-SAS

 Page 14 HUAWEI

- Disk Domain（磁盘域）
- Storage Pool（存储池）& Tier
- Disk Group（DG）
- LD（逻辑磁盘）
- Chunk（CK）
- Chunk Group（CKG）
- Extent
- Grain
- Volume & LUN

Disk Domain（硬盘域）

- Disk Domain即硬盘域，是一堆硬盘的组合（可以是整个系统所有硬盘），这些硬盘整合并预留热备容量后统一向存储池提供存储资源。





- OceanStor 18000系列存储系统可以一个或多个硬盘域
- 一个硬盘域上可以创建多个存储池（Storage Pool）
- 一个硬盘域的硬盘可以选择SSD、SAS、NL-SAS中的一种或者多种
- 不同硬盘域之间是完全隔离的，包括故障域、性能和存储资源等

Storage Pool（存储池） & Tier

- Storage Pool即存储池，是存放存储空间资源的容器，所有应用服务器使用的存储空间都来自于存储池。
- Tier即存储层级，存储池中性能类似的存储介质集合，用于管理不同性能的存储介质，以便为不同性能要求的应用提供不同存储空间。



 Page 16



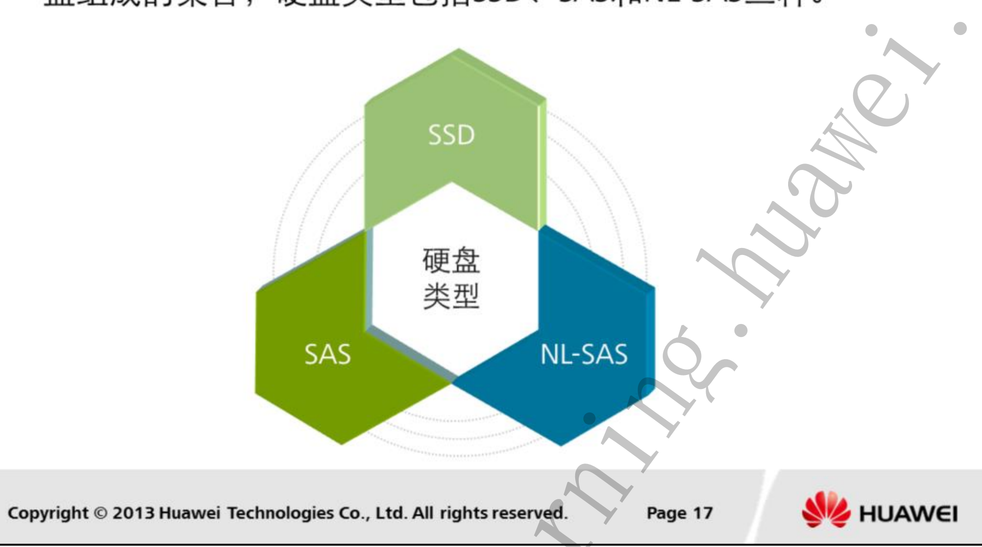
一个存储池基于指定的一个硬盘域创建，可以从该硬盘域上动态的分配Chunk（CK）资源，并按照每个存储层级（Tier）的“RAID策略”组成Chunk Group（CKG）向应用提供具有RAID保护的存储资源。

创建存储池可以指定该存储池从硬盘域上划分的存储层级（Tier）类型以及该类型的“RAID策略”和“容量”。

OceanStor 18000系列存储系统支持RAID5、RAID6和RAID10。

容量层由大容量的NL-SAS盘组成，RAID策略建议使用双重校验方式的RAID6。

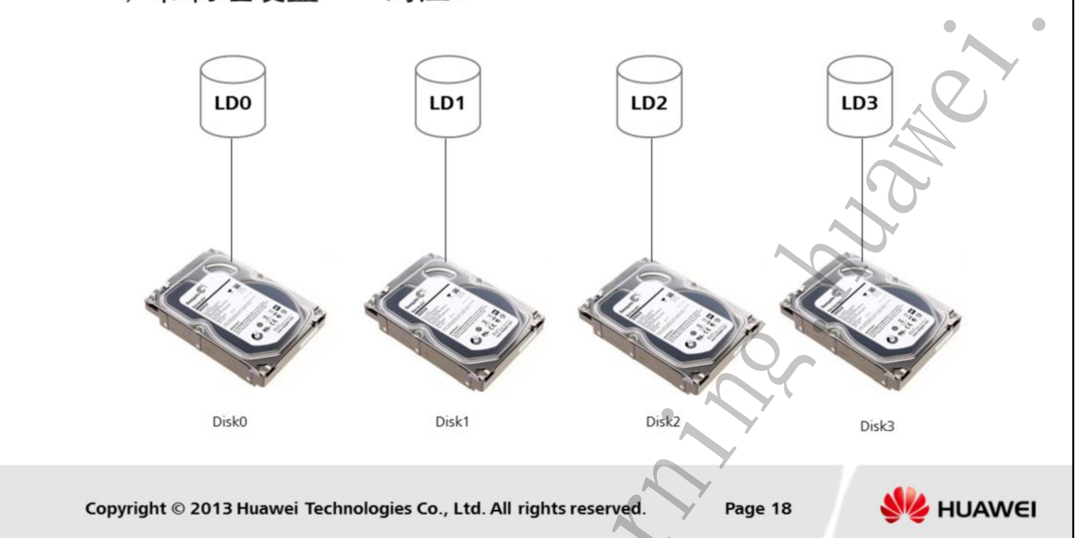
- RAID5：4D+1P，8D+1P
- RAID6：4D+2P，8D+2P
- RAID10：系统自动选择2D +2D或4D+4D

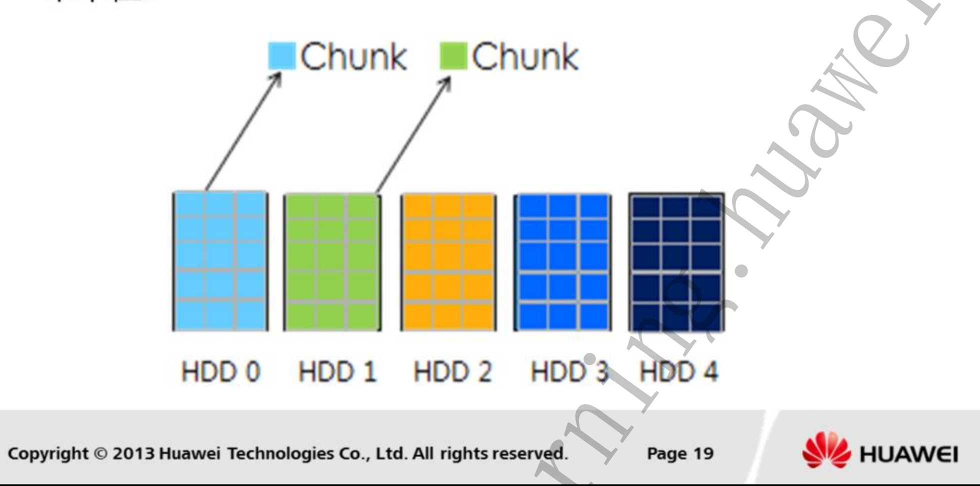


OceanStor 18000系列存储系统会在每个硬盘域内根据每种类型的硬盘数量自动划分为一个或多个Disk Group（DG）。

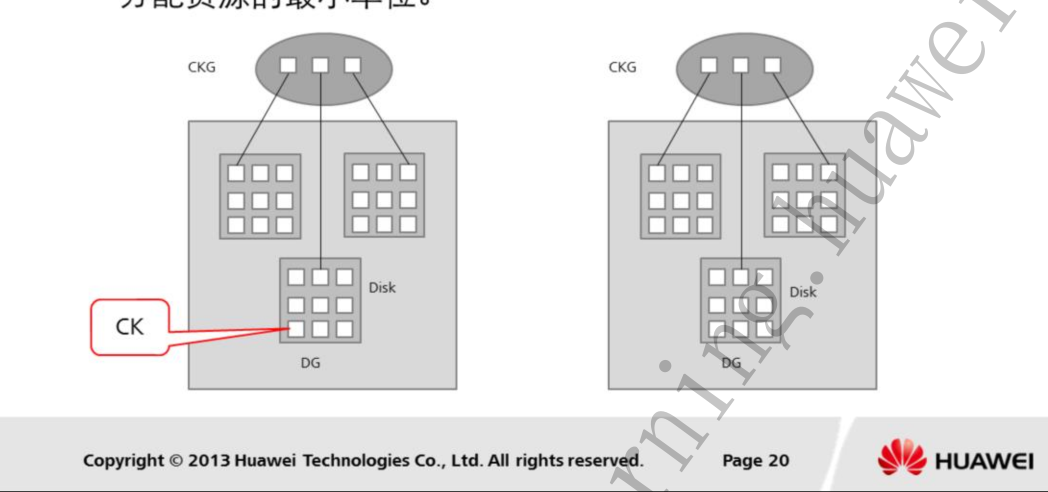
一个Disk Group（DG）只包含一种硬盘类型。

任意一个CKG的多个CK来自于同一个Disk Group（DG）的不同硬盘。


LD（逻辑磁盘）
Logical Drive（LD）即逻辑磁盘，是被存储系统所管理的硬盘，和物理硬盘一一对应。
LD0
LD1
LD2
LD3
Disk0
Disk1
Disk2
Disk3
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 18
HUAWEI



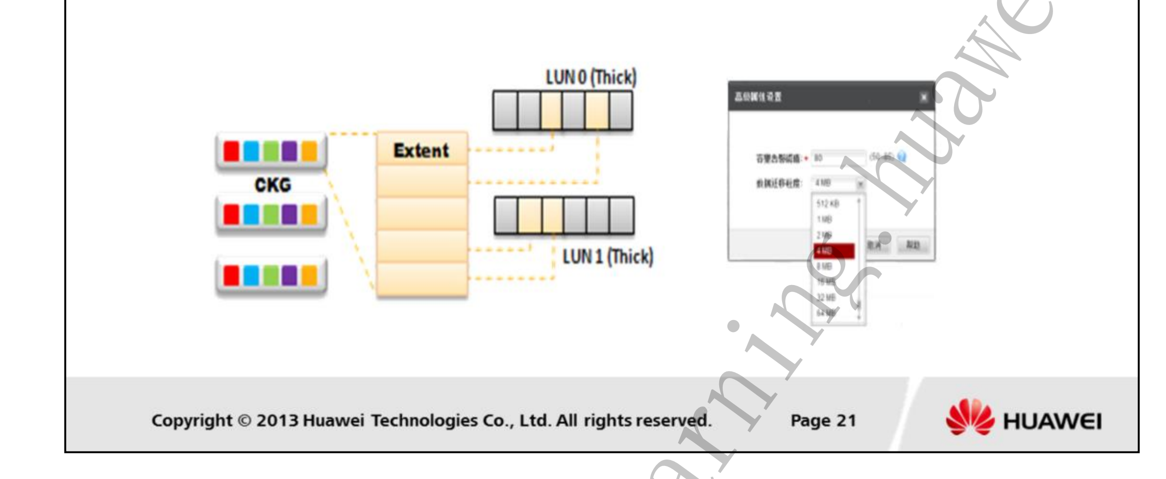
64MB的块大小是系统在切分物理空间的时候固定的大小，不能进行更改。



一个CKG中的CK均来自于同一个DG中的硬盘，CKG具有RAID属性（RAID属性实际配置在Tier上），CK和CKG均属于系统内部对象，由OceanStor 18000系统存储系统自动完成配置，对外不体现。

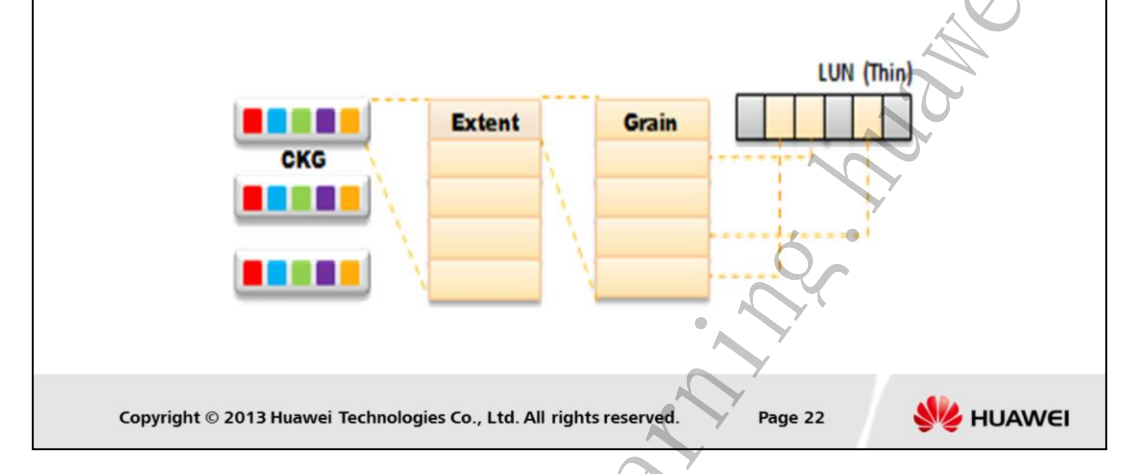
Extent

- Extent是在CKG基础上划分的固定大小的逻辑存储空间，大小可调，是热点数据统计和迁移的最小单元（数据迁移粒度），也是存储池中申请空间、释放空间的最小单位。



Page 21
HUAWEI

一个Extent归属于一个Volume或一个LUN，Extent大小在创建存储池时可以进行设置，创建之后不可更改，不同存储池的Extent大小可以不同，但同一存储池中的Extent大小是统一的



Thin LUN以Grain为单位映射到LUN，对于Thick LUN，没有该对象。

T系列存储V2的Grain粒度默认为32KB（可通过CLI指定范围为8KB-256KB），OceanStor 18000的Grain粒度固定为64KB。

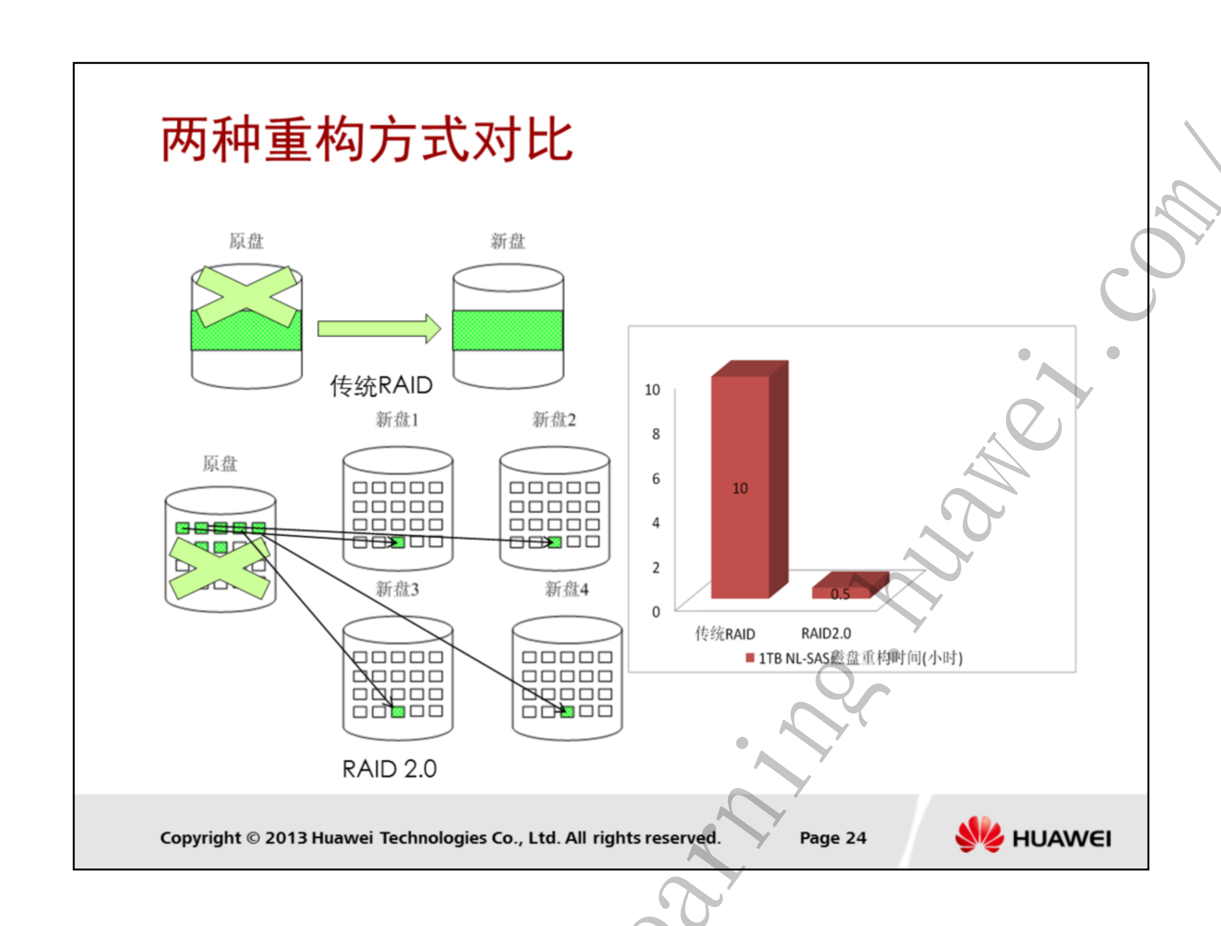
Volume & LUN

- Volume即卷，是存储系统内部管理对象。
- LUN是可以直接映射给主机读写的存储单元，是Volume对象的对外体现。

server
LUN
Volume
Storage

HUAWEI

一个Volume对象用于组织同一个LUN的所有Extent、Grain逻辑存储单元，可动态申请释放Extent来增加或者减少Volume实际占用的空间。



在传统RAID的重构中，故障盘的数据只能向一个热备盘上重构写。

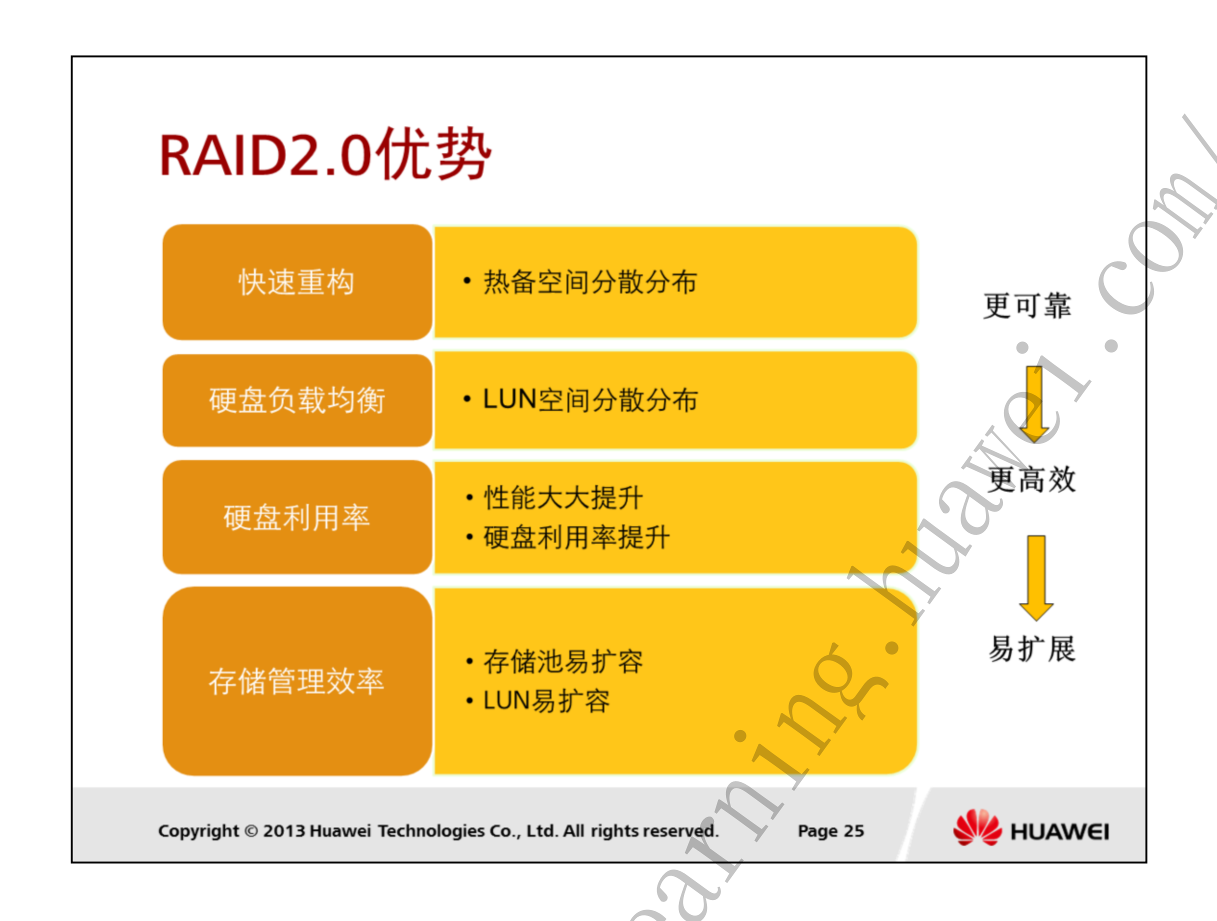
在RAID2.0的重构中，由于热备空间是分散在多个盘上的，故障盘上的数据可以重构到多个盘上，避免了对单热备盘的写瓶颈，因此重构速度很快。

另外，RAID2.0的重估是按Chunk重构的，重构时只重构LUN中分配了业务数据的Chunk，比传统RAID需要重构所有LUN空间的方式，减少了不必要的重构读写，也能提高整体重构速度。

那么，对重构速度的提升有多少呢?

1TB的NL-SAS盘的重构，使用传统RAID的方式需要10小时，而使用RAID2.0技术，只需要0.5个小时，也就是说，重构速度是原来的20倍。

重构速度的提高，减少了重构时间，因此减少了重构过程中对主机业务性能的影响时间，也减少了重构过程中双盘失效导致数据丢失的风险。



RAID2.0的最大优势是快速重构，但不仅于此。

硬盘负载均衡：LUN的数据被均匀分散分布到阵列内所有的硬盘上，可以防止局部硬盘过热，提升可靠性。在参与业务读写过程中，阵列内硬盘参与度高，提升系统响应速度。

最大化硬盘资源利用率：性能上，LUN基于资源池创建，不再受限于RAID组硬盘数量，LUN的随机读写性能可得到大大提升；容量上，资源池中的硬盘数量不受限于RAID级别，免除传统RAID环境下有些RAID组空间利用率高而有些RAID组空间利用率低的状况，并借助智能精简配置，提升硬盘的容量利用率。

提升存储管理效率：基于RAID2.0技术，我们无需花费过多的时间做存储预规划，只需简单地将多个硬盘组合成存储池，设置存储池的分层策略，从存储池划分LUN即可；当需要扩容存储池，只需插入新的硬盘，系统会自动的调整数据分布，让数据均衡的分布到各个硬盘上；当需要扩容LUN时，只需输入想要扩容的LUN大小，系统会自动从存储池中划分所需的空间，并自动调整LUN的数据分布，使得LUN数据更加均衡的分布到所有的硬盘上。

RAID2.0技术使得存储系统更可靠，更高效，易扩展，它是Smart tier、Smart Thin等特性的基础。首先我们介绍Smart Tier特性。



目录



SmartTier 原理

层级	硬盘类型	应用特点	数据特点
高性能层	SSD硬盘	适合随机读取存储请求密度高的业务负载。	最活跃数据：存储在或迁移至高性能层硬盘且读性能得到很大提高的“繁忙”数据。
性能层	SAS硬盘	适合存储请求密度中的业务负载	热数据：存储在或迁移至性能层硬盘的较活跃数据。
容量层	NL-SAS硬盘	适合存储请求密度低的业务负载	冷数据：存储在或迁移至容量层硬盘的“空闲”数据，且数据迁移后，其现有性能不会受到影响。

Page 27



首先，我们来详细看一下这个特性的原理。

SmartTier首先将不同存储介质按它们的性能高低和容量成本划分为三个存储层，由高至低分别是:

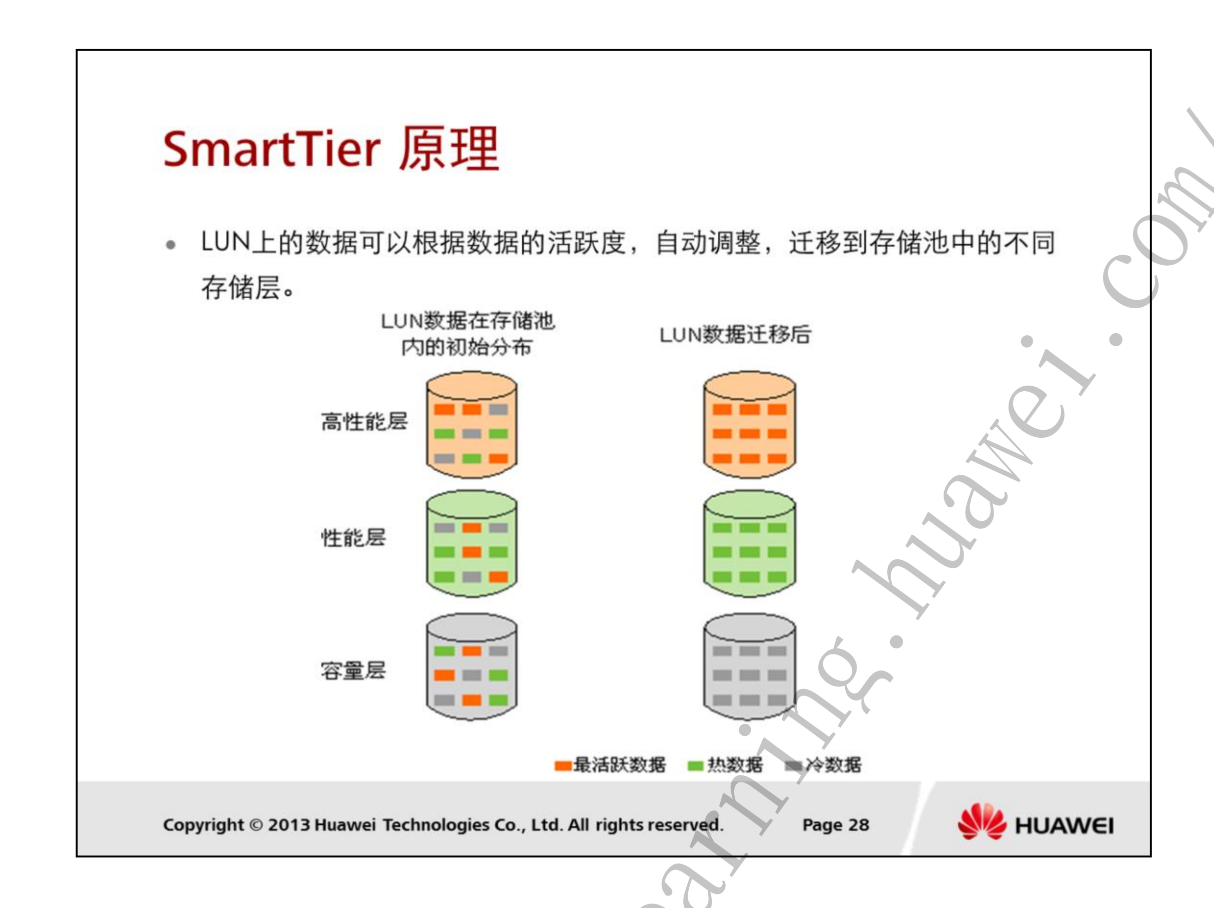
0级存储高性能层，统一使用响应速度最快的SSD硬盘，适合随机读取存储请求密度高的业务负载。

1级为性能层，使用速度和容量成本居中的SAS盘，适合存储请求密度中的业务负载;

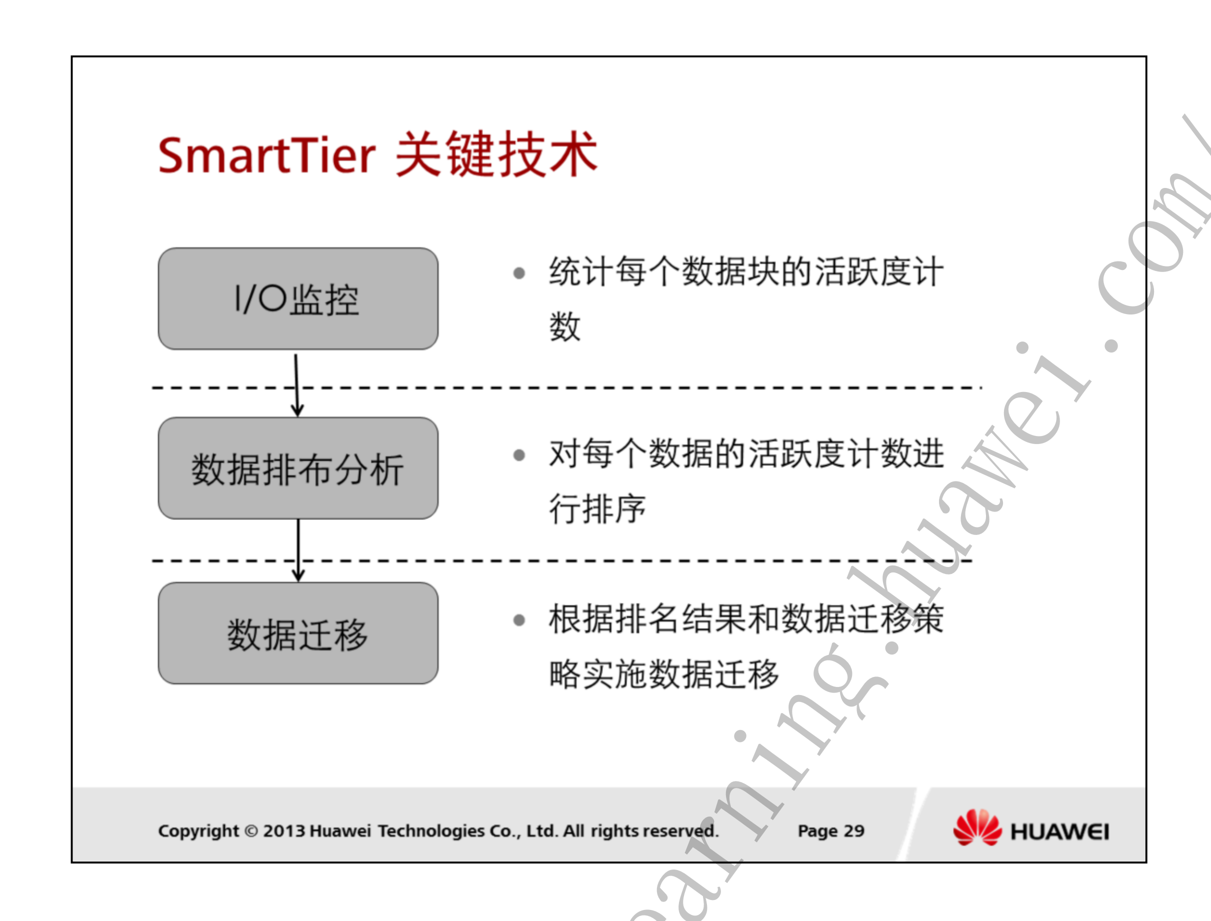
2级容量层，使用容量成本低而响应速度慢的NL－SAS盘，适合存储请求密度低的业务负载。

注意：每个存储层分别使用统一的硬盘类型。

根据各层的特点，在合适的配置下，Smar Tier能够自动识别数据的活跃程度，并将最活跃的数据存放或迁移到高性能层，热数据存放或迁移到性能层，冷数据存放或迁移到容量层。



Smart Tier的原理就是在存储池中创建LUN后，随着业务的读写，LUN上的数据可以根据数据的活跃度，自动调整，迁移到存储池中的不同存储层。此外，管理员也可以根据自己对数据活跃程度的预估，手动指定将数据优先存放于那一层。



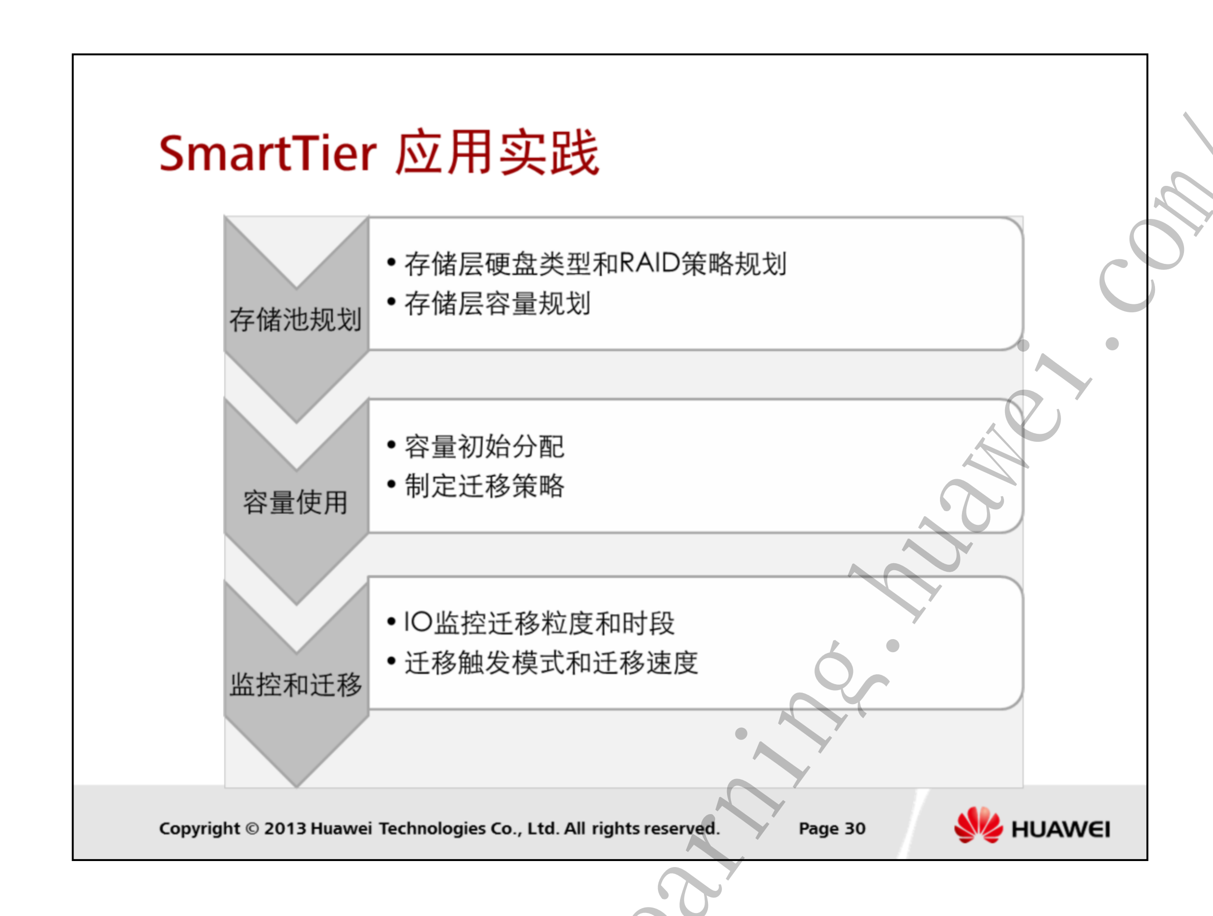
为了实现Smart Tier，使得数据能够根据活跃度自动在各层之间迁移，需要经历以下三个阶段。

首先，通过I/O监控，来统计每个数据块的活跃度计数

这里的数据块，是划分LUN数据大小的粒度，即extent，它是在存储池中申请空间、释放空间、迁移数据的最小单位，所以又称为"数据迁移粒度"。

然后，通过数据排布分析，来对每个数据块的活跃度计数进行排序

最后，就可以根据根据数据排布分析的排名结果和管理员指定的数据迁移策略实施数据迁移



作为一个系统管理员，如何根据自己对要放在存储系统上的实际业务的理解和要求，使得Smart Tier的特性发挥最佳的应用效果呢？

Smart Tier涉及以下三个方面的规划和使用：

- 首先要规划存储池各层的硬盘类型和相应的RAID策略，以及各层要规划的容量；
- 然后是存储池容量空间的使用，包括LUN的容量的初始分配和制定迁移策略；
- 最后是IO监控迁移粒度和时段的设置，以及迁移触发模式和迁移速度

SmartTier 应用实践

- 硬盘类型和RAID组策略

存储层	硬盘类型	RAID级别	读性能	写性能	硬盘利用率
0级 高性能层	SSD	RAID10	较高	较高	盘利用率为50%。
1级 性能层	SAS	RAID5	较高	较高	4D+1P（推荐）：硬盘利用率为4/5。 8D+1P：硬盘利用率约为8/9。
2级 容量层	NL-SAS	RAID6	中	中	4D+2P（推荐）：硬盘利用率约为4/6。 8D+2P：硬盘利用率约为8/10。

 Page 31 

首先，是存储池的规划。各存储层的硬盘类型，RAID组策略的配置要考虑到这一层的性能、可靠性以及成本的目标。

推荐配置可参考上图。

我们推荐在高性能层设置RAID 10，因为RAID 10具有较快的读写性能，这也与高性能层SSD硬盘的特点相适应。在性能层推荐设置RAID 5，在容量层设置RAID 6。因为通常RAID 5顺序写性能比RAID 10 更好，而读性能低于RAID 10。RAID 6具有双重数据校验功能，适合容量较大的NL SAS硬盘。

在存储池的每一层中，RAID组中数据空间和校验空间的比例对于Smart tier的性能表现也有所影响；根据理论分析和实验数据，对于RAID5，我们推荐采用4D＋1P的配置，对于RAID6，我们推荐4D+2P的配置。而RAID10的数据空间和校验空间之间是完全镜像且相等的，这个是固定的，因此无需配置。

存储池热备空间的策略有三种，分为高、低、无三种。分别是每6块和每12块硬盘对应一块硬盘的容量作为热备空间，以及不设置热备空间。在创建存储池时，需要留出足够的硬盘空间来做热备空间。

需要注意的是：RAID组级别，硬盘类型在创建存储池或扩容存储层时设置，且设置后不可更改；热备策略设置后可以修改。

SmartTier 应用实践

- 存储池热备策略

热备级别	描述
高	存储层每6块硬盘，使用一块硬盘的容量作为热备空间。
低	存储层每12块硬盘，使用一块硬盘的容量作为热备空间。
无	不设置热备空间

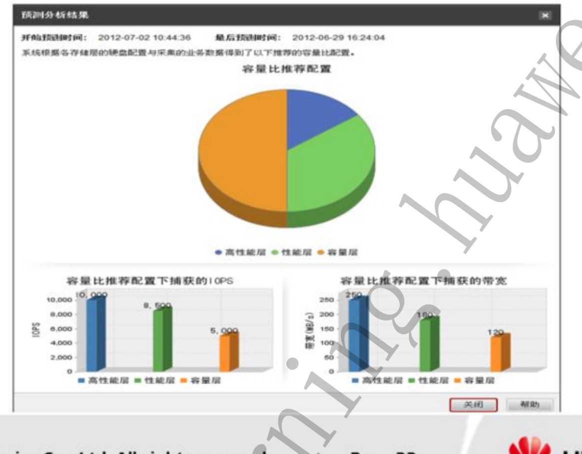


SmartTier 应用实践

存储层容量规划

理论上，存储层容量策略是各个存储层的存储容量比例为1：1：1。

为防止单一应用程序消耗所有SSD硬盘资源，限制该应用程序所在存储池的高性能层存储容量。





 Page 33 HUAWEI

在创建存储池时，对于每个存储池各层配备多大的容量？

当然，最理想的存储层容量策略是各个存储层的存储容量比例为1：1：1，，因为该策略允许将各个存储层100%的数据升级或降级至其他任一存储层。但在实际应用中该策略通常并不合适，由于某些原因可能需要限制访问某个特定存储层。

比如，为防止单一应用程序消耗所有SSD硬盘资源，限制该应用程序所在存储池的高性能层存储容量可能是合适的。这种情况下，推荐使用包含较小比例的高性能层存储容量策略。

另外，OceanStor 18000智能存储系统还可以根据业务的实际情况，监控数据的冷热比例，给出最佳容量比推荐配置，帮助用户动态调整各存储层的容量。

上图是一个预测分析的实例。我们可以从饼图中看到容量比的推荐配置，并在柱状图中看到在这样的推荐配置下能获得的IOPS和带宽。

SmartTier 应用实践

- LUN数据迁移策略
  - 不迁移
  - 向高性能层迁移
  - 向低性能层迁移
  - 自动迁移（默认策略）

LUN容量的初始分配

自动分配	分配顺序是Tier1、Tier2、Tier0
指定Tier分配	管理员指定分配的级别

LUN应用场景	推荐容量初始分配策略	推荐数据迁移策略
对性能要求较高的LUN	优先从高性能层分配或优先从性能层分配	向高性能层迁移
对性能要求不敏感的LUN	优先从容量层分配	向低性能层迁移
大规模数据写入LUN	优先从高性能层分配	自动迁移或不迁移

系统默认设置是自动分配和自动迁移。

LUN的初始容量分配策略要根据LUN的应用场景，结合LUN的数据迁移策略共同考虑。

 Page 34



LUN容量的初始分配策略是在创建LUN时设置，表示要创建的这个LUN的初始容量希望分配在那个存储层，设置后不可更改。

配置策略有两种方式:自动分配和指定Tier分配。

自动分配是由系统自动决定为LUN初始分配哪个存储级别的空间。分配顺序是Tier1、Tier2、Tier0，即性能层，容量层，高性能层。只有在前面一层Tier分配不到空间时才到下一层Tier分配。

指定Tier分配，由系统管理员根据业务的特征来指定分配在哪个级别，如“优先从高性能层分配”、“优先从性能层分配”或“优先从容量层分配”。只有在被指定的级别没有空间时，才分配到下一层。

LUN的数据迁移策略用于指定LUN数据的迁移方向;

- 不迁移：LUN的数据在不做任何迁移动作；
- 自动迁移：SmartTier根据存储池中数据块的活跃度排名结果进行数据迁移。自动迁移依据I/O监控得出的分析数据进行迁移，因此必须启用I/O监控并设置业务监控时段后，自动迁移才能生效。
- 当选择向高性能层迁移和向低性能层迁移后，不论这些LUN是什么活跃级别，SmartTier会将他们的数据块优先迁移至相应的层级去。

系统默认设置是自动分配和自动迁移。

SmartTier 应用实践

LUN容量的初始分配

自动分配	分配顺序是Tier1、Tier2、Tier0
指定Tier分配	管理员指定分配的级别

LUN数据迁移策略

- 不迁移
- 向高性能层迁移
- 向低性能层迁移
- 自动迁移（默认策略）

- 系统默认设置是自动分配和自动迁移。
- LUN的初始容量分配策略要根据LUN的应用场景，结合LUN的数据迁移策略共同考虑。

LUN应用场景	推荐容量初始分配策略	推荐数据迁移策略
对性能要求较高的LUN	优先从高性能层分配或优先从性能层分配	向高性能层迁移
对性能要求不敏感的LUN	优先从容量层分配	向低性能层迁移
大规模数据写入LUN	优先从高性能层分配	自动迁移或不迁移

 Page 35



在实际使用中，LUN的初始容量分配策略要根据LUN的应用场景，结合LUN的数据迁移策略共同考虑。对性能要求较高的LUN，适合优先从高性能层分配策略或优先从性能层分配策略，同时其数据迁移策略适合向高性能层迁移。 对性能要求不敏感的LUN，适合优先从容量层分配策略，同时其数据迁移策略适合优先向低性能层迁移。

大规模数据写入存储池，适合优先从高性能层分配策略，同时其数据迁移策略适合自动迁移或不迁移。这样做的原因是如果运行该业务的LUN使用自动分配策略，部分数据最开始将分配至性能层和容量层。此时，SmartTier还未对新写入数据进行I/O监控和数据排布分析，无法分辨数据的活跃度。因此部分最活跃数据可能将分配至性能层和容量层，而不能优先使用高性能层。

LUN的数据迁移策略设置后可在使用过程中根据需要更改。

SmartTier 应用实践

- 业务监控时段
  - 推荐进行默认的全天候I/O监控。
  - 应用程序处于非活动状态或非主业务运行的时段应排除。
- IO监控，即数据迁移粒度
  - 流媒体、视频监控，推荐：4MB
- 迁移触发模式
  - 在新增加一个存储层，或修改LUN数据的存储位置且希望立即生效时，推荐“手动迁移”
  - “定时迁移”，推荐设置在业务空闲时段，如每天凌晨1点～4点
- 迁移速度
  - 业务繁忙时，推荐“低”
  - 业务空闲时，推荐“中”、“高”

 Page 36 

为进行指定存储池SmartTier数据块的排布分析，需要对存储池各层数据块的IO个数，平均IO大小等统计指标进行监控，并指定监控的日期和时间范围。

推荐进行默认的全天候进行I/O监控。

如果应用程序很长时间将处于非活动状态，或者应用程序在某时间段产生的I/O不是用户的主业务（例如后台备份、数据同步）则业务监控时段应排除该时间段。

数据迁移粒度，即extent，是存储池分配LUN空间的粒度，也是IO监控，数据迁移的粒度。extent越小，存储资源的使用效率越高，数据迁移的有效性越高。但是，对于流媒体业务、视频监控业务等大IO业务，粒度大则读写和迁移速度快，此时推荐采用的最佳做法是使用默认值4MB。

迁移触发模式分为手动迁移模式和定时迁移模式。手动方式可在任一时刻设置时间段进行数据迁移，且设置后立即生效，推荐在新增加一个存储层时，或业务变化需要修改LUN的存储位置且希望立即生效时使用；定时方式只能在预先设置的时间段进行数据迁移，推荐采用的最佳做法是将数据迁移时间设置在业务空闲状态时间段，例如每天凌晨1点～4点。迁移触发模式设置后可以在使用中更改。

迁移速度分为高中低三种。无论使用哪一种，SmartTier都会根据当前业务负载，在上限范围内动态调整迁移速率，确保数据迁移不会对当前业务造成明显影响。

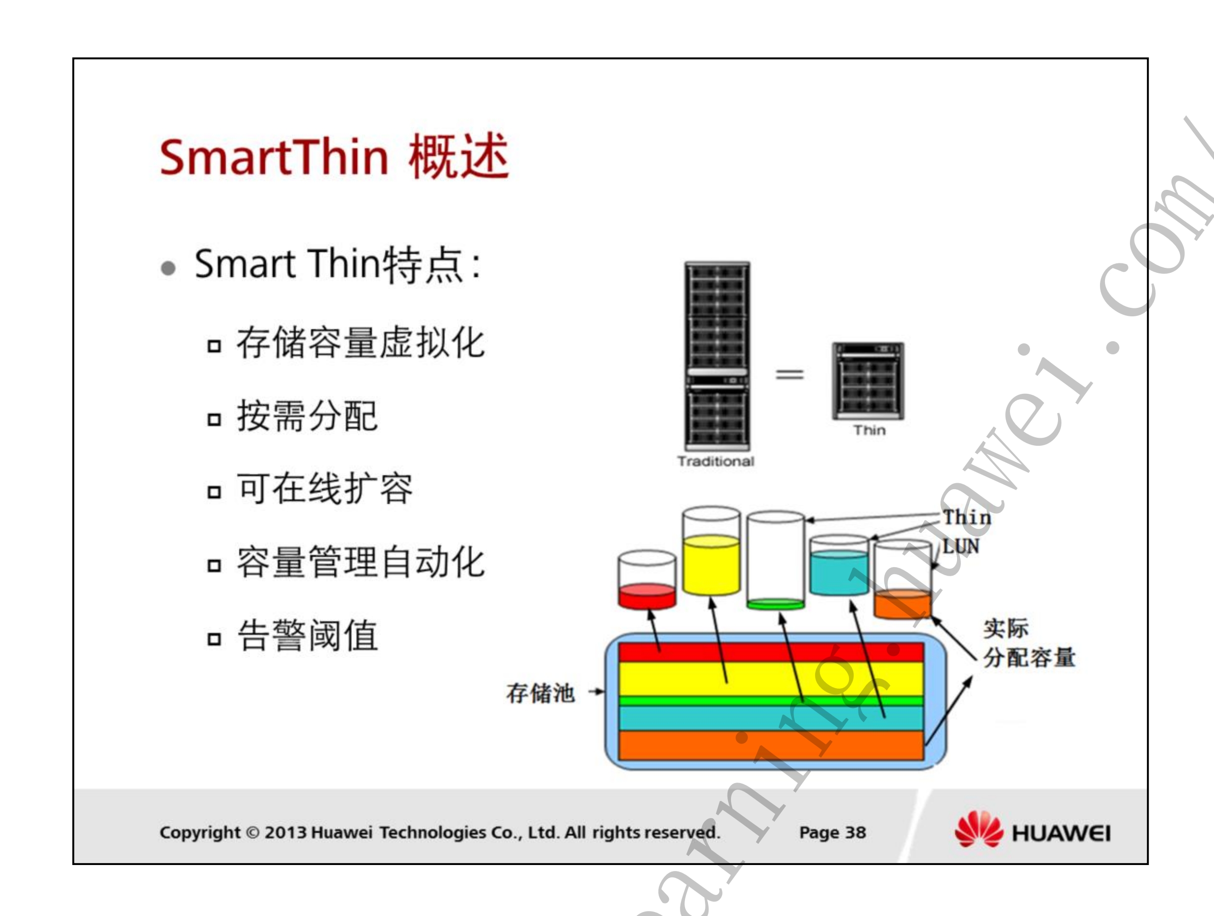
推荐做法是：当业务处于繁忙状态时，推荐采用的最佳做法是设置较保守的数据迁移速率（“低”）。当业务处于空闲状态时，可以将数据迁移速率设置为更激进的级别（“中”或“高”），这将允许SmartTier更快速、动态的针对业务负载的细微变化做出调整。

通过以上的配置，SmartTier动态分级存储特性，能自动将不同活跃度的数据和不同特点的存储介质动态匹配起来，降低了用户的成本，提高了系统的整体性能。



目录





Smart Thin，是一种存储容量的虚拟化特性。如左图所示，对于主机来说Thin LUN和传统LUN是没有区别的。但是对存储系统来说，Thin LUN只需从存储池中分配实际使用到的空间即可。这种按需分配的技术可以以很少的初期投入，满足用户不断增长的存储容量需求，使得存储系统具有很高的利用率和可扩展性。它能够在线扩容，扩容时不必对数据进行迁移或备份，避免了数据迁移带来的风险并节约了数据备份带来的成本。可以为用户提供自动化的容量管理功能，用户不必再费心于为不同的业务配置不同的容量，它的容量竞争机制使得各种业务通过竞争获取所需容量，最终达到存储容量的最优化配置。我们还可以为Thin LUN所在的存储池设置一个告警阈值，当应用程序实际使用的存储池的容量接近该阈值时，系统会上报一条告警，提示我们为存储扩充容量。

实现Thin LUN空间按需分配的关键技术有两个，一个是写时空间分配的方式来分配空间，另一个是通过读写重定向来支持数据的读写。下面我们详细来介绍一下这两种技术。

SmartThin 关键技术

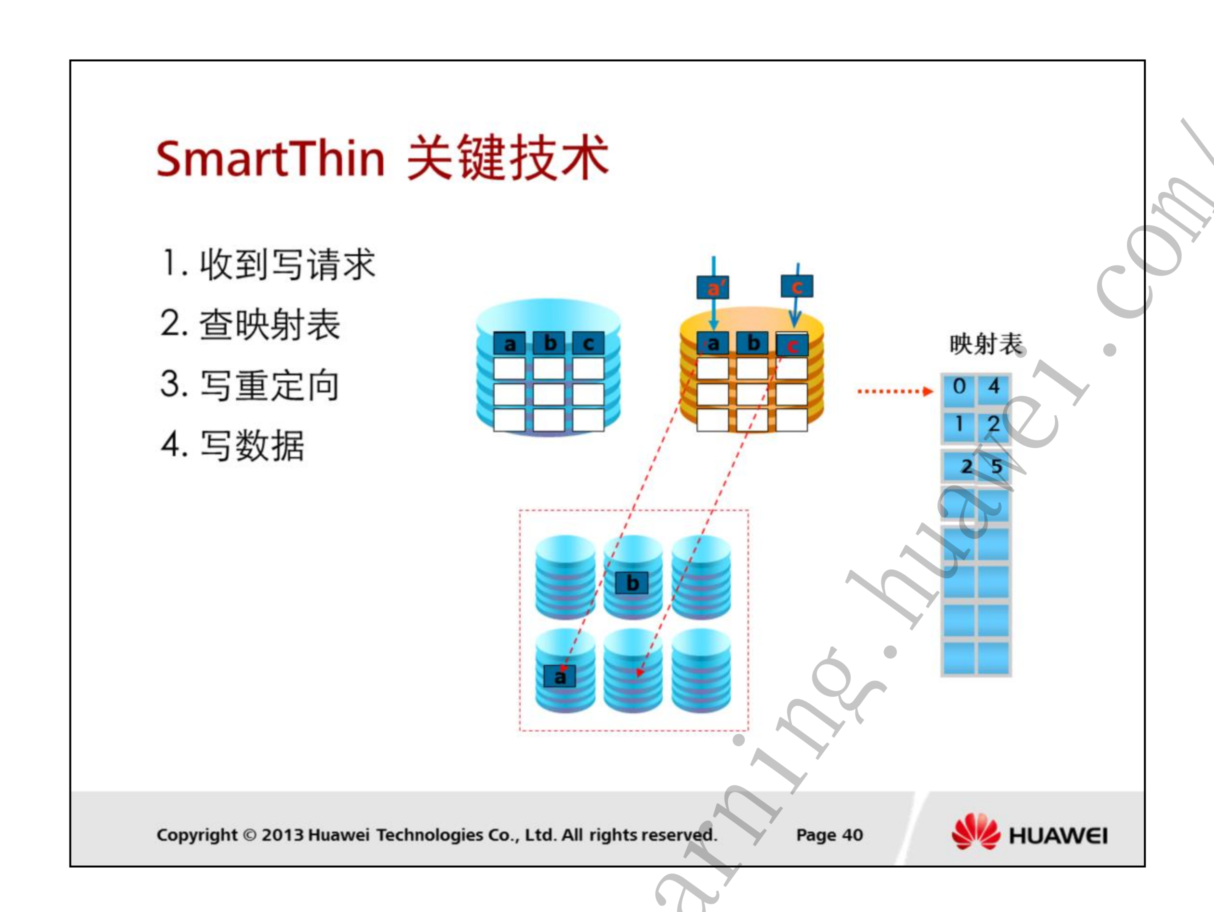
- 写时空间分配： Capacity-on-write
- 读写重定向： Redirect-on-time

Thin LUN
存储实际数据
物理存储空间（存储池）
RAID5
空间分配
主机
写操作
映射
用时再分配

Page 39
HUAWEI

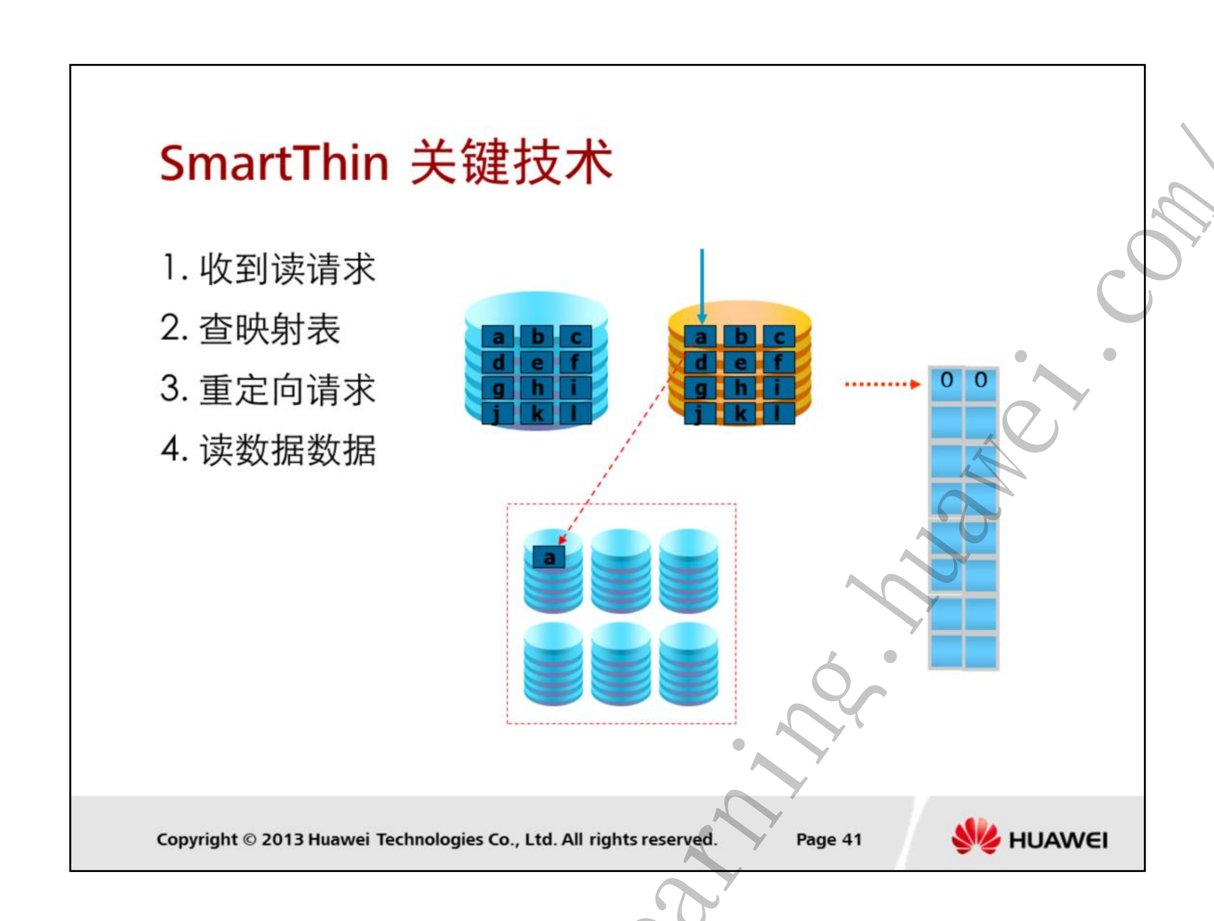
写时空间分配技术是指对Thin LUN的写IO请求触发存储池的实际空间分配，而未写到的位置等到用时再分配。写时空间分配技术，是一种动态分配空间的技术，写数据时分配的实际存储区域是不确定的，需要专门的映射表来记录数据的逻辑地址和实际存储位置的对应关系。

读写重定向技术是指对Thin LUN进行读写时需要根据映射表进行重定向。



映射表的映射项的左项为逻辑地址，作为查找键值。右项记录资源块的地址，形象的说就是“指针” 。

Thin LUN收到主机写请求时，先查映射表，如果未记录，则分配存储池中的物理空间，进行写，并记录映射表；如果映射表中已记录了目标地址在存储池中的物理地址，则重定向到该物理地址进行覆盖写。



Thin LUN收到主机读请求时，先查映射表，根据映射表中已记录的目标地址对应的在存储池中的物理地址，然后重定向到该物理地址进行读。

SmartThin 应用实践

- 适用的场合

业务类型	分析	举例
对业务连续性要求较高的系统核心业务	在线对系统进行扩容，不会中断业务	银行票据交易系统
应用系统数据增长速度无法准确评估的业务	按需分配物理存储空间，避免浪费	E-mail邮箱服务、网盘服务等
多种业务系统混杂并且对容量需求不一的业务	让不同业务去竞争物理存储空间，实现物理存储空间的优化配置。	运营商服务等

- 不适用的场合

业务类型	分析	举例
对I/O性能要求很高的场合	读写重定向，对性能有一定影响	在线交易

 Page 42



Smart Thin适用于对业务连续性要求较高的系统核心业务，因为它可以支持在线对系统进行扩容，不会中断业务。例如：银行票据交易系统。也适用于应用系统数据增长速度无法准确评估的业务，因为它可以按需分配物理存储空间，避免浪费，例如：E-mail邮箱服务、网盘服务等。

对于多种业务系统混杂并且对存储需求不一的业务也同样适用，因为它可以让不同业务去竞争物理存储空间，实现物理存储空间的优化配置，例如：运营商服务等。但是由于Smarthin采用了读写重定向技术，对性能有一定影响，因此不适用于对I/O性能要求很高的场合，如在线交易。



目录

 Page 43



Smart QoS概述

- 背景：
  - 不同应用程序之间由于业务模型和I/O特征不同相互影响，导致存储系统整体性能受到影响；
  - 不同应用程序相互争抢系统带宽和IOPS资源，关键业务性能无法得到保证。
- 需求：
  - 保证关键型应用程序的性能；
  - 保证高级别用户的性能。
- Smart QoS原理：
  - 该特性允许用户根据应用程序的一系列特征（IOPS或带宽），对每一种应用程序设置特定的性能目标；
  - 存储系统根据设定的性能目标，动态分配存储系统的资源来满足特定应用程序的服务级别要求，从而优先保证关键性应用程序服务级别的需求。

Page 44



随着存储系统容量持续增大，将多个应用程序部署在同一台存储设备上的需求也在逐步增加。将多个应用程序部署在同一台存储设备可以简化用户的存储系统架构，但是，不同的应用程序之间由于业务模型和I/O特征不同相互影响，导致存储系统整体性能受到影响。

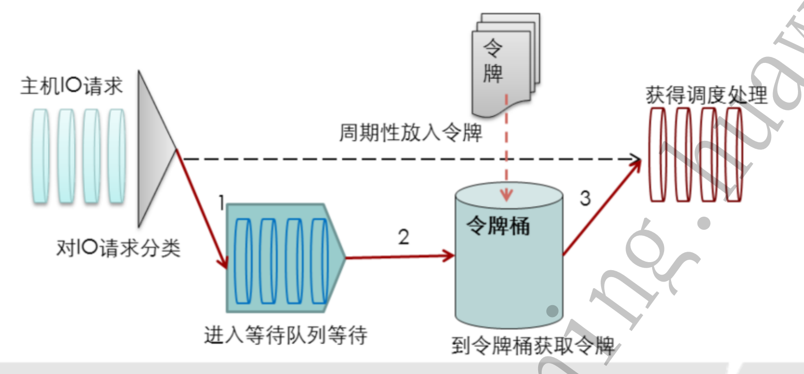
此外，出于投资成本的考虑，某些用户可能不会单独投资存储系统的建设。此时，可借助存储资源提供商提供的存储资源运行相关业务，减少总体投资且能保证业务的连续运行。但由于每个用户的业务类型及业务特征不同，不同的应用程序相互争抢系统带宽和IOPS资源，关键业务性能和高级别用户所需的存储资源可能无法得到满足。

为了使关键型应用程序和高级别用户的性能得到保证，OceanStor 18000存储系统为用户提供了Smart QoS功能。

Smart QoS的原理是：允许用户根据应用程序的一系列特征（IOPS或带宽），对每一种应用程序设置特定的性能目标；存储系统再根据设定的性能目标，动态分配存储系统的资源来满足特定应用程序的服务级别要求，从而优先保证关键性应用程序服务级别的需求。

SmartQoS 关键技术

- SmartQoS基于令牌桶原理
  - 用户每配置一个SmartQoS策略，系统会根据用户设置的性能目标生成一个令牌桶，按照用户配置的性能目标周期性向令牌桶中放入一定数量的令牌。
  - 每一个受这个SmartQoS 策略控制的I/O请求都必须从令牌桶中获得一个令牌才能得到处理；如果令牌桶中的令牌取空，则只能在等待队列中等待系统下一次放入令牌。

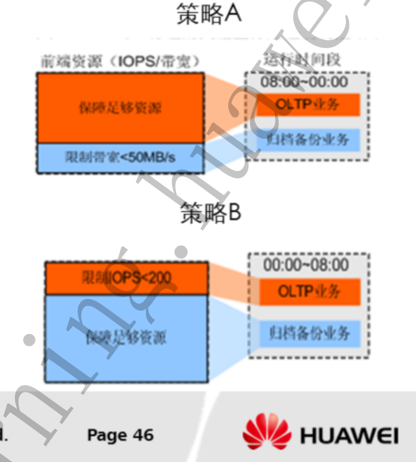




SmartQoS 应用实践

业务类型	I/O特征	主要运行时间段
在线交易业务	随机小I/O，通常以IOPS来衡量。	08:00至00:00
归档备份业务	顺序大I/O，通常以带宽来衡量。	00:00至08:00

- 这两种业务在各自对应的时间段内都需要保障足够的系统资源。
- 创建两个Smart QoS策略：
  - 策略A：
    - 在08:00至00:00时间段内限制备份归档业务的带宽（例如<50MB/s）。
  - 策略B：
    - 在00:00至08:00时间段内限制在线交易业务的IOPS（例如<200）。



 Page 46 HUAWEI

假设我们的系统中同时部署了On-Line Transaction Processing在线事务处理系统(在线交易) 和归档备份业务，应该如何配置Smart QoS策略呢?

首先要分析这两种业务的IO特性和主要运行时段。在线交易业务通常是随机小IO，以IOPS来衡量性能；主要运行时段早上八点，到晚上12点；归档备份业务通常是顺序大IO，以带宽来衡量性能；主要运行时段是晚上12点到早上8点前。

这两种业务在各自对应的时段内都要保障足够的系统资源。用户可通过SmartQoS特性，创建两个Smart QoS策略来达到业务需求。

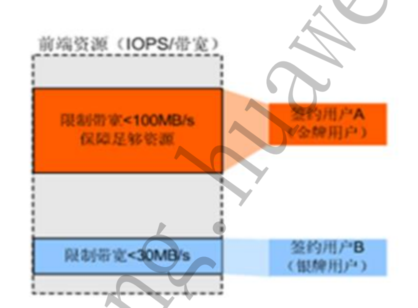
策略A：在08:00至00:00时间段内限制备份归档业务的带宽（例如<50MB/s），从而预留足够系统资源给在线交易业务。

策略B：在00:00至08:00时间段内限制在线交易业务的IOPS（例如<200），从而预留足够系统资源给备份归档业务。

SmartQoS 应用实践

用户分类	业务质量要求
签约用户A（金牌用户）	高
签约用户B（银牌用户）	中

- 要优先保证高级别用户的业务质量。
- 创建两个Smart QoS策略：
  - 策略A：
    - 限制金牌用户A的业务带宽（例如<100MB/s）。
  - 策略B：
    - 限制银牌用户B的业务带宽（例如<30MB/s）。



 Page 47



假设我们的系统中同时部署了金牌用户A和银牌用户B的业务时，应该如何配置Smart QoS策略呢？因为这两个用户属于不同级别，对业务质量的要求不同，因此在资源紧张的情况下，要优先保证高级别用户的业务质量。同样，我们需要创建两个SmartQoS策略。

策略A：限制金牌用户A的业务带宽（例如<100MB/s）。

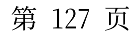
策略B：限制银牌用户B的业务带宽（例如<30MB/s）。

此带宽比金牌用户A小，从而预留足够系统资源给金牌用户A。



目录





Smart Partition概述

- 背景：
  - 随着存储系统的不断完善和发展，仅支撑单业务存储服务的存储系统正在被可以同时向数千个应用提供存储服务的存储系统所替代。但如何同时满足数千个不同应用的巨量并发，这给存储系统的服务质量保证能力提出了更高的要求。
- 需求：
  - 根据应用的优先级不同，保证各个级别的应用性能。
- Smart Partition原理：
  - SmartPartition特性以业务LUN为单位进行SmartPartition分区设定，每个SmartPartition分区的资源独立访问，互不干扰。

 Page 49 

同时，存储系统将保证各分区中应用的缓存容量，并根据实际情况自动调整不同分区的应用服务器侧和存储系统侧的I/O并发，从而保证位于该分区的业务LUN的性能。

SmartPartition 关键技术

- SmartPartition特性通过隔离不同业务所需的缓存资源，保证关键业务的服务质量。通过SmartPartition特性，用户将引擎内的缓存资源划分为2个或2个以上的分区（少于8个），每个分区的读写缓存容量根据用户业务LUN的I/O情况进行设置，提高缓存读写速度的同时大大缩短存储系统的应用响应时间。
- SmartPartition特性通过周期统计每个SmartPartition分区的所有主机并发量，然后通过用户针对每个LUN设置的读写缓存容量以及每个LUN的优先级，自动匹配主机并发，保证SmartPartition分区资源的最大化使用的同时，保证关键业务的服务质量。

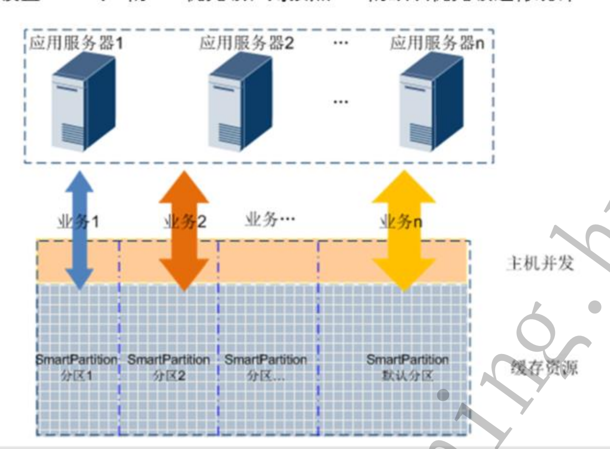
 Page 50



在存储系统中，影响某业务的服务质量的主要因素包括缓存容量和主机并发，其中缓存容量表示该业务当前可以占用存储系统的缓存资源，主机并发表示存储系统针对该业务允许的主机I/O并发处理量。

SmartPartition 关键技术

- 在主机并发中，LUN的优先级的判断有两种方式：
  - 如果针对LUN设置了SmartQoS的LUN优先级，则按照SmartQoS优先级进行统计。
  - 如果未设置SmartQoS的LUN优先级，则按照LUN的默认优先级进行统计。



HUAWEI

SmartPartition 应用实践

- 在VDI（Virtual Desktop Infrastructure）场景中，每个用户的应用及应用服务质量的要求各不相同，数据中心的难题是如何完全满足每个用户的要求，同时还可充分利用现有资源。
- 通过SmartPartition特性可以为不同的签约用户所使用的LUN创建不同的缓存分区容量，在资源有限的情况下可以优先保证高级别用户的业务质量要求。
- 例如，数据中心的存储资源同时为多个用户提供存储资源，其中用户A和用户B的特征如表所示

用户分类	业务质量要求
用户A（金牌用户）	高
用户B（银牌用户）	中

- 存储资源提供商通过SmartPartition特性可以创建两个缓存分区，两个缓存分区设置不同的读写策略。

 Page 52



- SmartPartition策略A：为用户A使用的LUN创建SmartPartition分区1（例如读缓存10GB，写缓存15GB），该读写缓存容量保证用户A的应用正常运行且读写性能优异。
- SmartPartition策略B：为用户B使用的LUN创建SmartPartition分区2（例如读缓存5GB，写缓存8GB），此缓存容量比用户A小，从而预留足够系统资源给用户A。



目录



Hyper Snap 概述

- 定义：
  - Hyper Snap，即快照，是指在不中断正常业的情况下、源数据在某一时间点的数据的完全可用的拷贝，该拷贝是源数据在拷贝时间点的静态映像。
- 特点
  - 快照可以对存储设备上的数据灵活和频繁地生成多个恢复点，在需要时可以快速地恢复数据。
  - 快照可瞬间生成，不影响源LUN业务，可以方便地作为备份和归档的数据源。
- 应用
  - 数据备份与恢复、持续数据保护、结合其他软件实现备份和容灾，并应用于报表生成、数据测试、数据分析等数据处理场景。

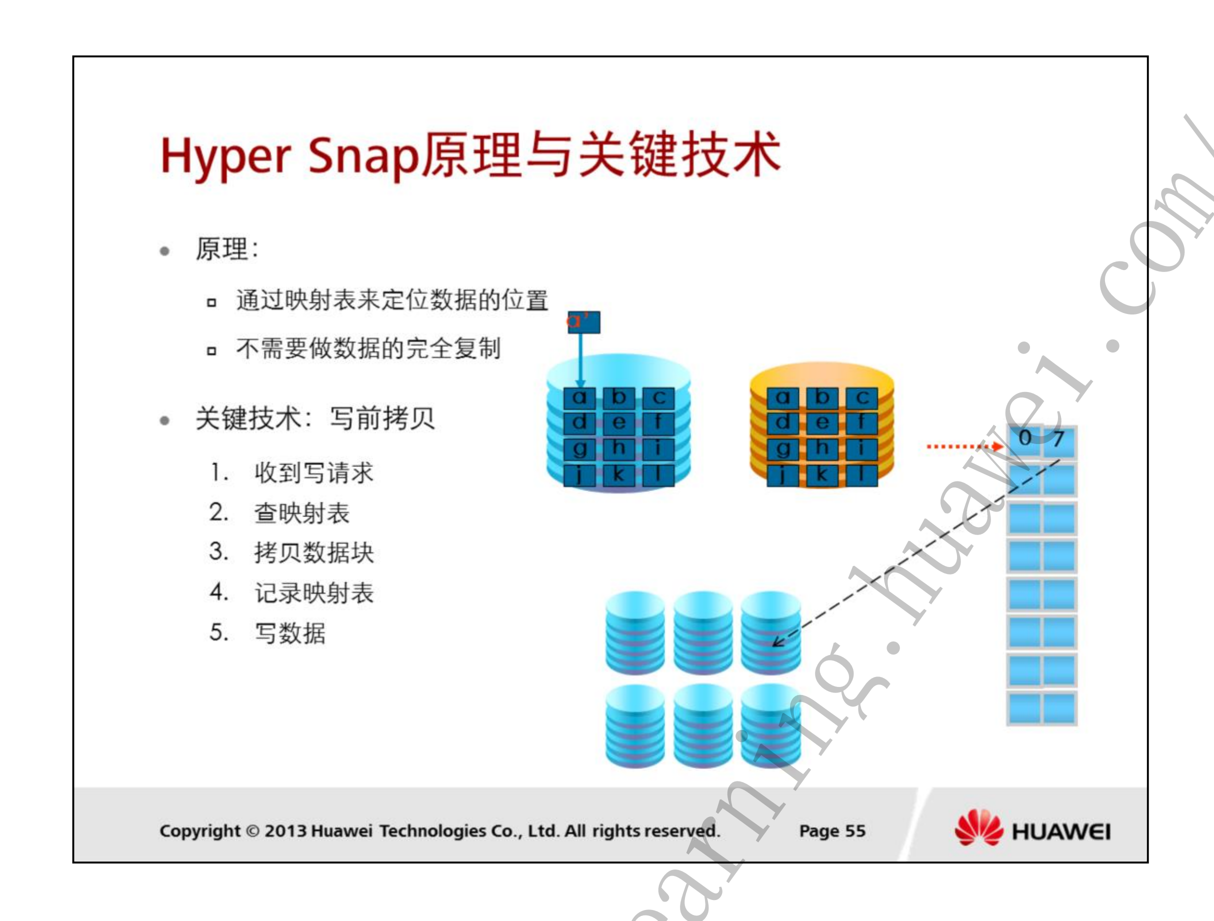
 Page 54



Hyper Snap，即快照，是指在不中断正常业的情况下、源数据在某一时间点的数据的完全可用的拷贝，该拷贝是源数据在拷贝时间点的静态映像。静态的意思时，在快照激活后，会保持与激活时的源数据完全一致，而不会随着源数据的改变而改变。

其特点是：快照可以对存储设备上的数据灵活和频繁地生成多个恢复点，在需要时可以快速地恢复数据。快照可瞬间生成，不影响源LUN业务，可以方便地作为备份和归档的数据源。

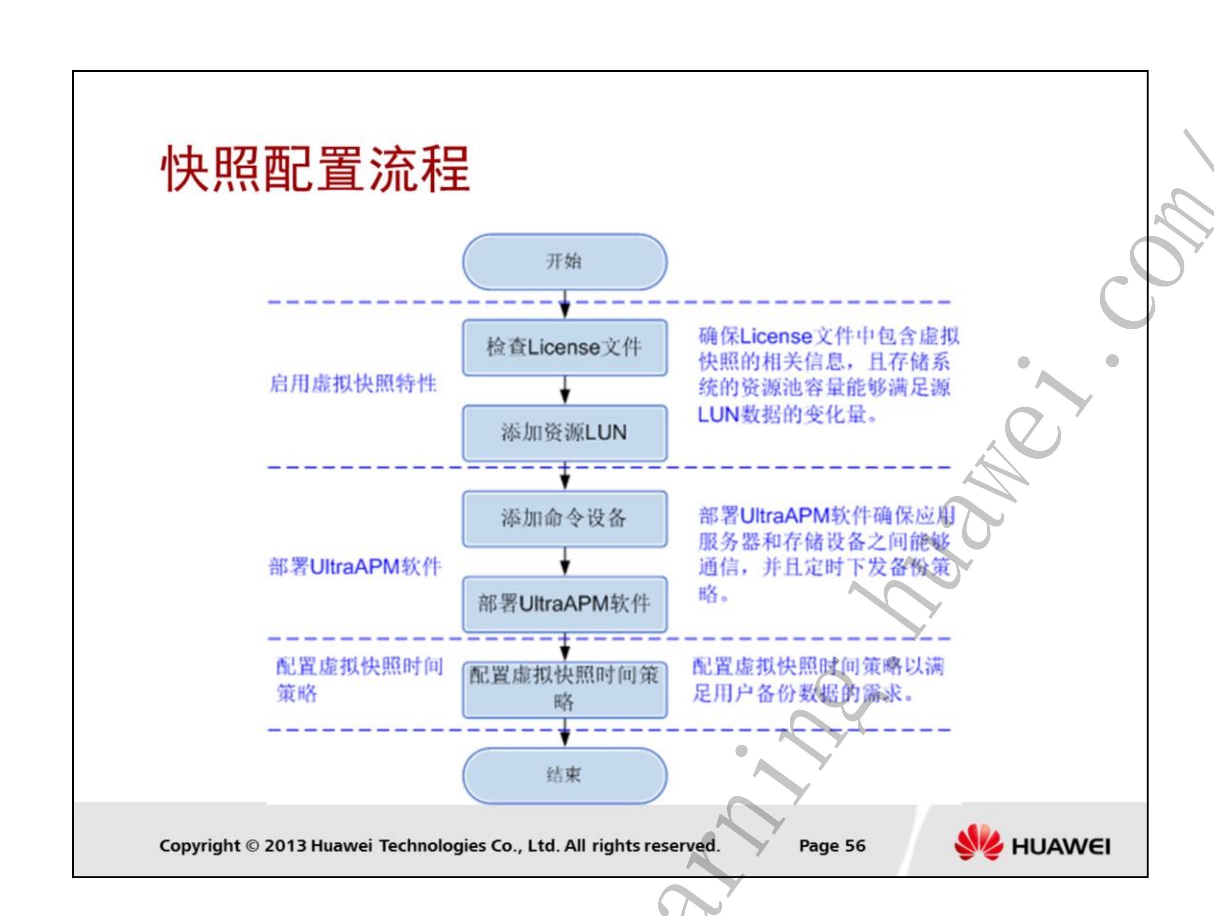
通过生成不同时间点的快照，或者同一时间点的多份快照，您可以实现快速的数据备份与恢复、持续数据保护、结合其他软件实现备份和容灾，并应用于报表生成、数据测试、数据分析等数据处理场景。



快照的原理是通过映射表来定位数据的位置，而不需要做数据的完全复制。映射表的左项是源地址，作为查找键值，右项为数据存放的实际位置。每一个快照都有一个映射表与之对应。

实现快照的关键技术是写前拷贝。快照激活时，它的映射表里没有项，这表示源LUN的数据跟快照激活时还没有发生变化，读快照LUN时，可以直接读相应的源LUN。当主机向源LUN写入新的数据时，系统先查映射表，如果没有记录有这个数据要写入的位置的表现，说明源LUN中，该数据对应位置还没有被修改过，这是第一次修改。因此需要将要被覆盖掉的原数据搬移到一个新的位置，并向映射表中增加一行，这一行的左项指向源LUN的数据位置，右项指向原数据的新位置。之后，再将主机下发的新数据实际写入。读快照LUN时，当发现映射表中对于要读到的源LUN位置有记录，就从映射表右项对应位置读取数据。否则就从源LUN上读取数据。

通过这种写前拷贝的技术，就能通过映射表来定位数据的位置，只有在源LUN要被修改时才做拷贝，而不需要做源LUN所有数据的完全复制。


快照配置流程
开始
检查License文件
添加资源LUN
启用虚拟快照特性
确保License文件中包含虚拟快照的相关信息，且存储系统的资源池容量能够满足源LUN数据的变化量。
添加命令设备
部署UltraAPM软件
部署UltraAPM软件
部署UltraAPM软件确保应用服务器和存储设备之间能够通信，并且定时下发备份策略。
配置虚拟快照时间策略
配置虚拟快照时间策略
配置虚拟快照时间策略以满足用户备份数据的需求。
结束
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 56
HUAWEI

快照重要概念

- 源LUN
- 资源池
- 资源LUN
- 映射表
- 快照LUN

 Page 57 

- 源LUN
    - 需要进行虚拟快照操作的源数据所在的LUN。
- 资源池
    - 在虚拟快照过程中，用于临时保存源LUN改变前原数据的存储空间。由一个或多个LUN组成。
- 资源LUN
    - 被添加到资源池中的一个或多个LUN。这些LUN的容量构成资源池容量。
- 映射表
    - 用于记录快照LUN中数据的存放地址。映射表主要由两部分组成：数据在源LUN中地址及数据在资源池的地址。正常情况下，映射表保存在存储系统的内存中。但是，当存储设备运行于故障模式（如电源故障、镜像链路故障）时，存储系统将实时把映射表写入到保险箱硬盘中。
- 快照LUN
    - 对源LUN创建虚拟快照后，逻辑上生成的数据副本，由映射表地址指向的数据组成的虚拟集合。

快照相关操作

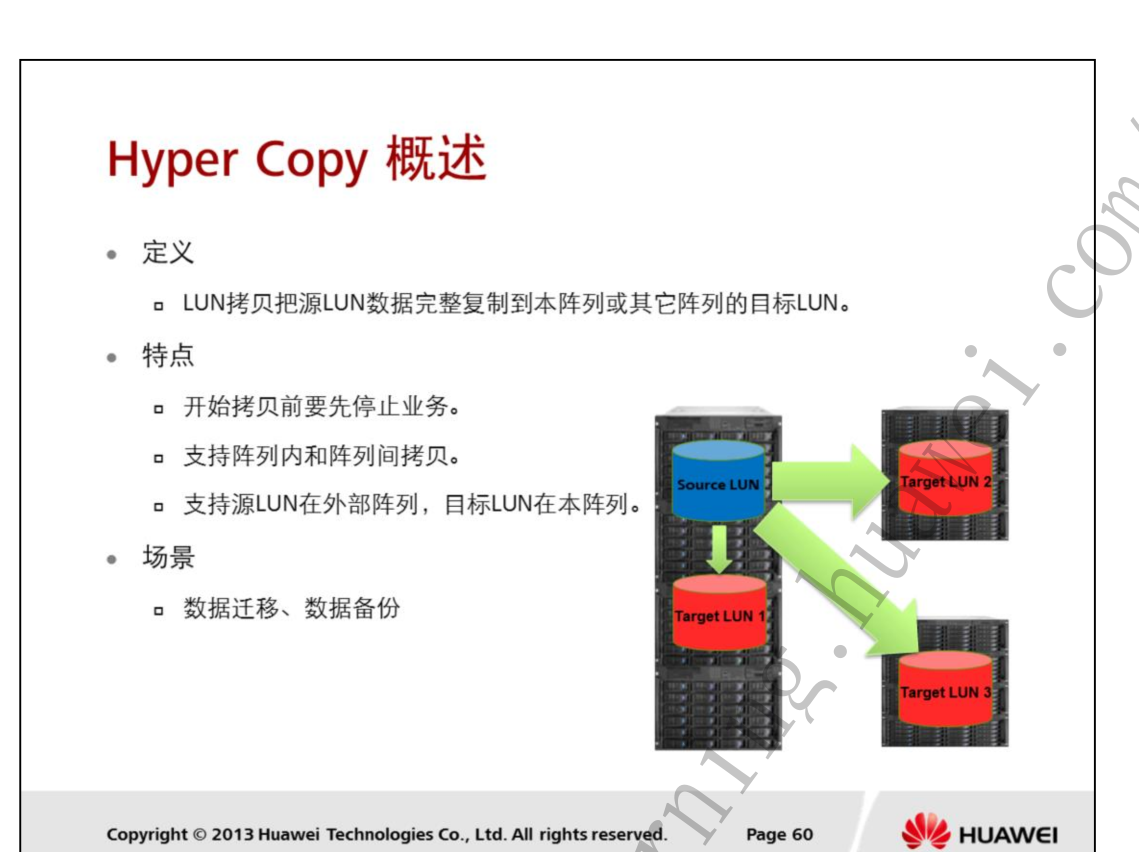
相关操作	操作结果
激活	当用户激活一个快照任务时，存储系统在激活时间点生成一份快照，得到激活时间点源LUN的一致性副本。
停止	当用户停止一个快照任务时，存储系统释放占用的资源LUN的空间，快照LUN不可用。
重建	当用户重建一个快照任务时，存储系统会在重建时间点重新激活快照，快照LUN中的数据变成源LUN在重建时间点的一致性副本。
回滚	当用户回滚一个快照任务时，存储系统用快照LUN的数据覆盖源LUN的数据，将源LUN的数据置于快照时间点的状态。





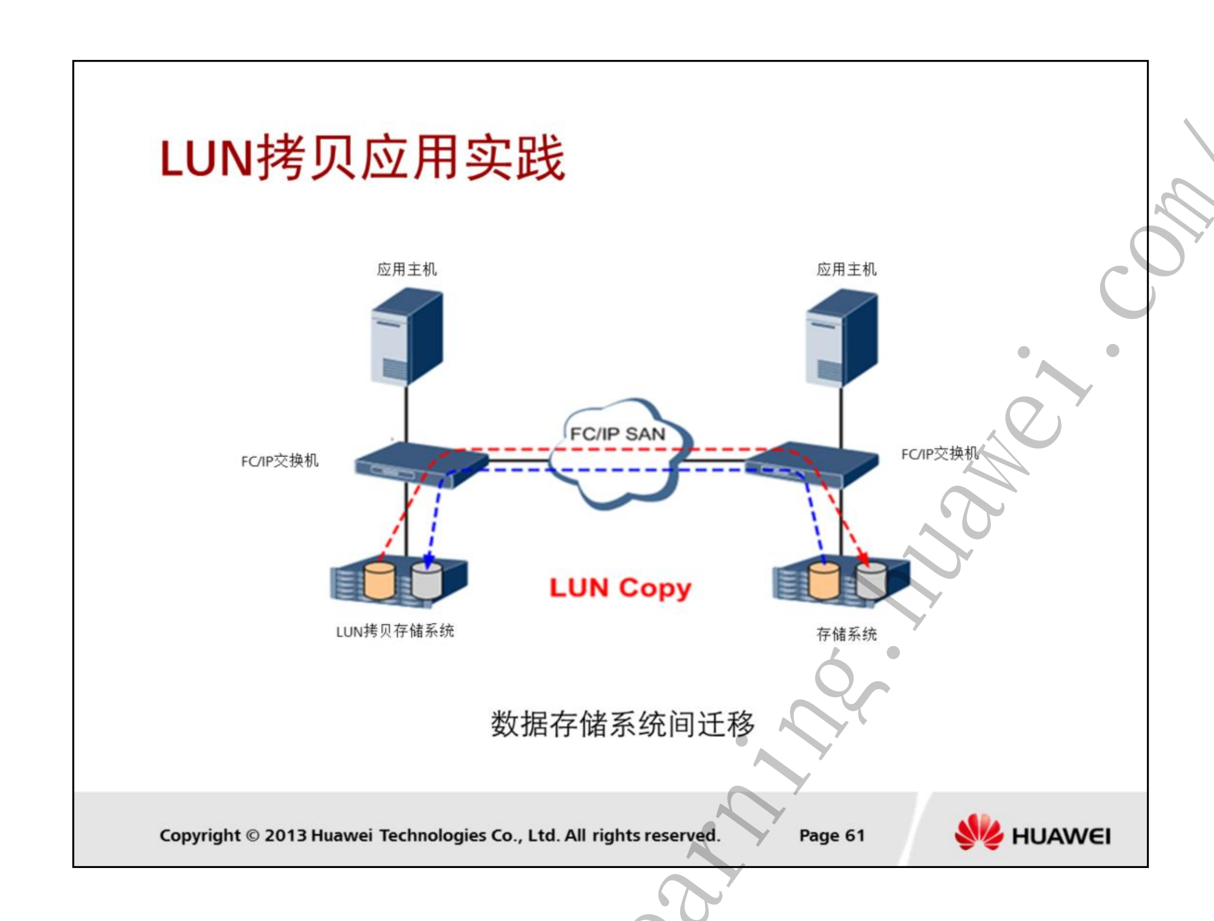
目录



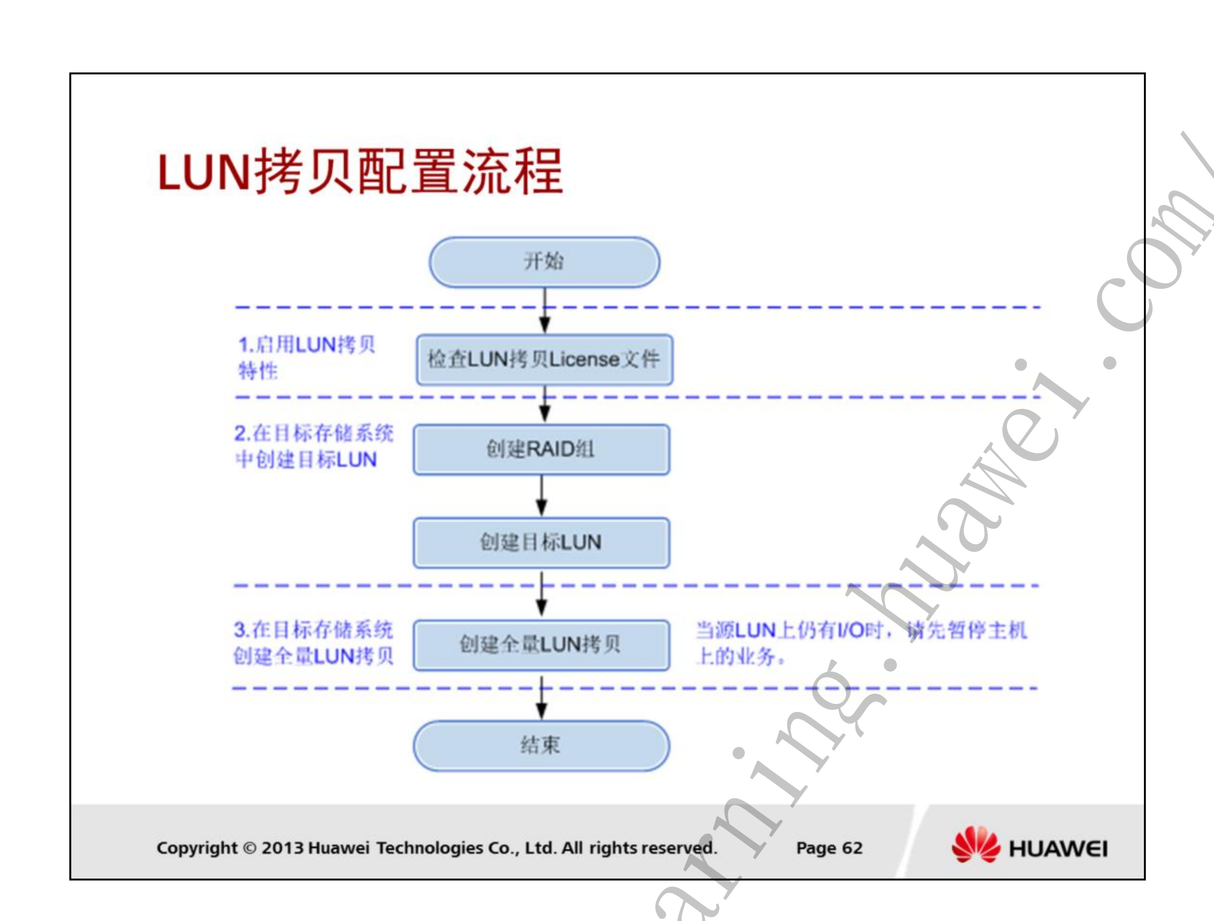


LUN拷贝的能把源LUN数据完整复制到本阵列或其它阵列的目标LUN。

- 它的特点有：
  - 为了保证拷贝过程中，目标LUN和源LUN的数据一致性，开始拷贝前要先停止源LUN的业务。
  - LUN拷贝支持阵列内和阵列间拷贝。
  - 它还支持源LUN在外部阵列，目标LUN在本阵列之间的拷贝。
- LUN拷贝的使用场景有：
  - 数据迁移和数据备份。



在存储系统升级、业务迁移、新业务试运营等情况下，需要对生产数据进行迁移。此时，通过全量LUN拷贝，可以很方便进行存储系统内或存储系统之间的数据迁移。


LUN拷贝配置流程
开始
1.启用LUN拷贝特性
检查LUN拷贝License文件
2.在目标存储系统中创建目标LUN
创建RAID组
创建目标LUN
3.在目标存储系统创建全量LUN拷贝
创建全量LUN拷贝
当源LUN上仍有I/O时，请先暂停主机上的业务。
结束
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 62
HUAWEI

LUN拷贝重要概念

- 源LUN
- 目标LUN
- 备份窗口
- 内部快照
- 差异位图
- 进度位图

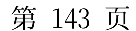
 Page 63 

- 源LUN
  - 源数据所在的LUN。
- 目标LUN
  - 拷贝后数据所在的LUN。
- 备份窗口
  - 在进行备份操作时，主机业务需要暂停的时间。
- 内部快照
  - 在增量LUN拷贝启动时会被激活，快速地为源LUN产生完全可用的数据副本，供目标LUN进行数据拷贝。
- 差异位图
  - 记录源、目标LUN数据的差异。
- 进度位图
  - 记录源、目标LUN差异数据同步的进度。



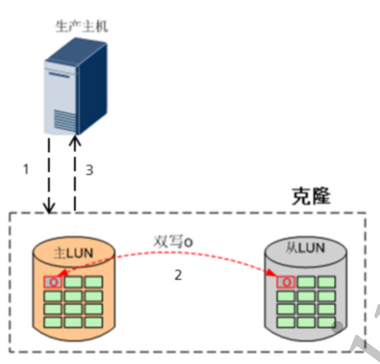
目录





克隆特性介绍

- 概述
  - 克隆是在不中断正常业务的前提下，在物理上生成主LUN在某个时间点的一份完整拷贝。
  - 一个克隆组支持一个主LUN对应八个从LUN，可以对同一个主LUN备份八份不同时间点的数据。



Page 65



克隆实现原理如上图：

1. 主机写主LUN
2. 克隆同时写主LUN和从LUN
3. 主从LUN都写成功后，克隆返回主机写成功
4. 使用从LUN时，将从LUN与主LUN分裂，从LUN数据与分裂时主LUN数据一致。

克隆应用实践

- 克隆比较典型的应用之一是升级演练。
- 使用克隆所生成的数据副本可以并行开展应用测试、数据分析等 业务活动。

应用测试主机 生产主机 数据分析主机

I/O 主LUN I/O

从LUN 从LUN ...... 从LUN

 Page 66 HUAWEI

注意，在使用数据副本时需要将副本与主LUN分裂才能使用。分裂时，从LUN的数据与此时主LUN的数据完全一致。



目录



远程复制 特性介绍

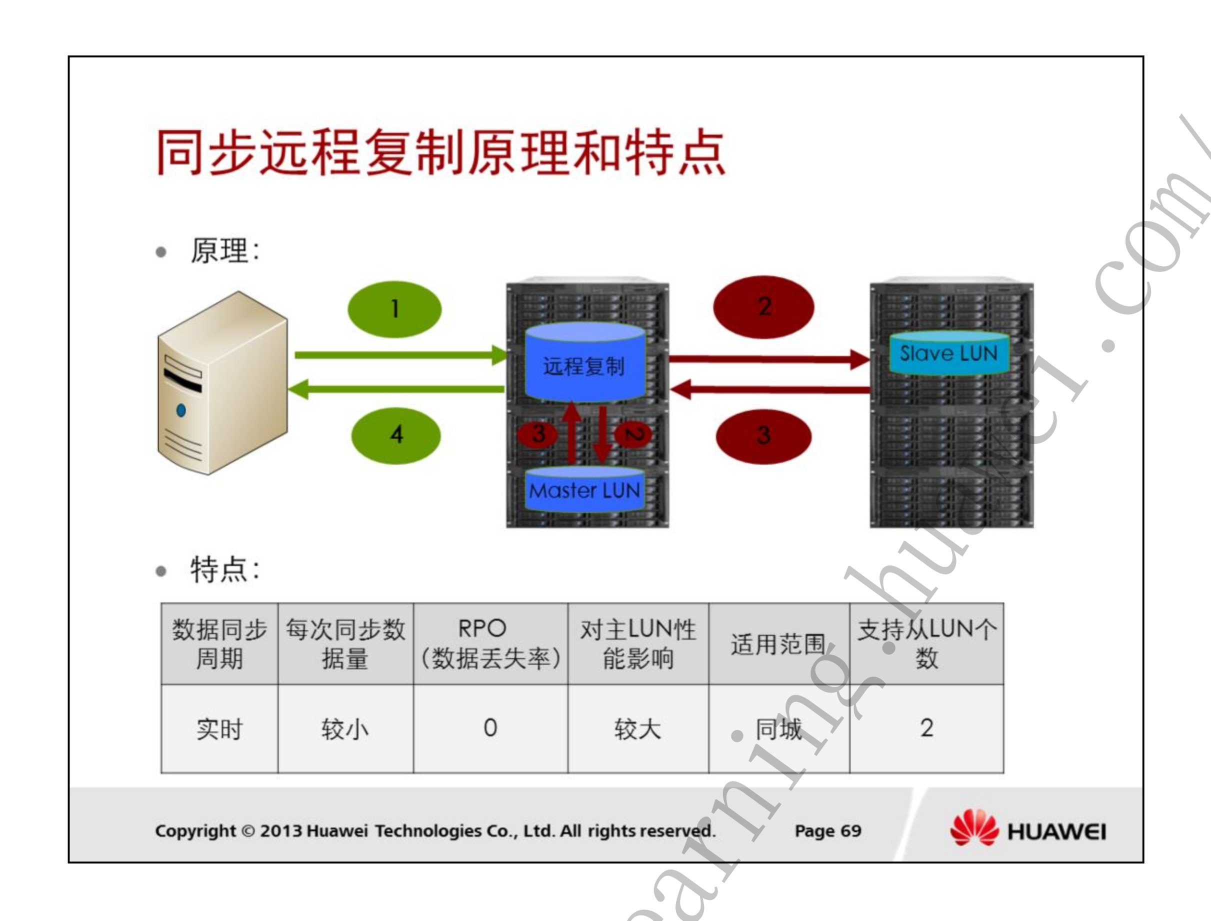
- 定义：
  - 远程复制(Remote Replication)是数据镜像技术的一种，通过在两个或多个站点维护若干个数据副本，实现数据的远端备份和恢复、持续的业务数据支撑、数据的迁移和分发、数据的灾难恢复。
- 分类：
  - 同步远程复制
  - 异步远程复制

 Page 68



为了解决远程复制时主机业务系统性能与主站点阵列和从站点阵列（后简称主从站点）数据一致性之间的矛盾，远程复制分为以下两种复制模式，同步远程复制和异步远程复制。

- 同步远程复制：
  - 实时地同步数据，最大限度保证数据的一致性，以减少灾难发生时的数据丢失量。
- 异步远程复制：
  - 周期性地同步数据，最大限度减少由于数据远程传输的时延而造成的业务性能下降。



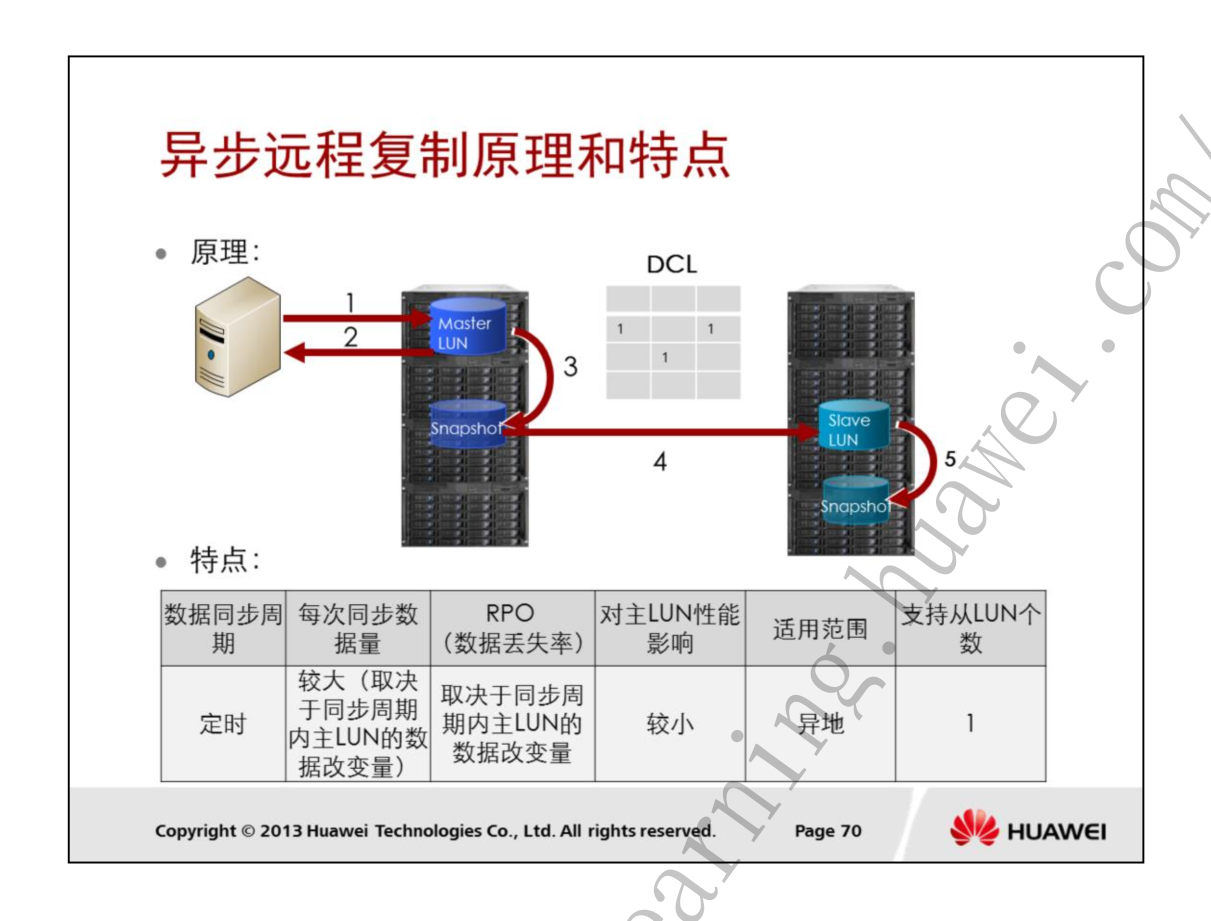
我们首先来了解在同步远程复制模式中，主机IO的处理流程。

主机发送写I/O至阵列；主站点阵列将写I/O的数据写入主LUN并发送写I/O至从LUN；从站点将写I/O的数据写入从LUN；待主、从LUN写I/O均成功后，阵列的主LUN向主机返回写I/O成功。

同步远程复制的特点如下：

由于主机每次写都要同时写主从LUN，因此其同步周期是实时的，每次同步的数量量较小。当主LUN发生故障时，从LUN的数据由于总是与主LUN实时同步的，数据的丢失率，即RPO为0。但是，由于主机的一次写IO需要实时同步到从LUN后才能返回写成功，增加了IO时延因此对主LUN的性能影响较大；随着主从阵列之间的距离加大，这个时延也加大。因此同步远程复制的适用范围是同城的。

存储系统的同步远程复制可以支持一个主LUN有两个从LUN，但是这个两个从LUN必须是两个不同的从阵列。



数据同步周期	每次同步数据量	RPO（数据丢失率）	对主LUN性能影响	适用范围	支持从LUN个数
定时	较大（取决于同步周期内主LUN的数据改变量）	取决于同步周期内主LUN的数据改变量	较小	异地	1

现在来了解在异步远程复制模式中，主机IO的处理流程。

首先，在创建异步远程复制关系后，主从站点的阵列要分别为主LUN和从LUN创建一个快照。在主机将写I/O至发送给主站点阵列后；主站点阵列将写I/O的数据写入主LUN，就向主机返回写I/O成功。同时，主站点将主从LUN的差异记录在数据差异表DCL中。在启动主从LUN的同步时，激活主从LUN的快照，将差异数据同步到从LUN。

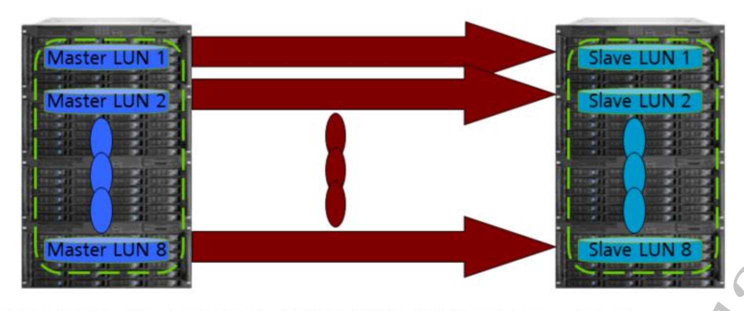
这里，主LUN快照的作用是保证同步过程中读取到的主LUN数据的一致性，从LUN快照可以备份从LUN在同步开始之前的数据，避免同步过程中的异常情况，从而导致从LUN数据的不可用。

异步远程复制的特点如下：

由于主机每次写只写主LUN，主从LUN的数据同步是周期性的，因此其同步周期是定时的，每次同步的数量量取决于同步周期内主LUN的数据改变量。主站点发生故障后，从LUN保留的是上一个周期同步的数据，本周期的数据还没有同步，将会丢失。因此，数据丢失率RPO也取决于同步周期内主LUN的数据改变量。但是，这种同步模式的好处是主机写IO时只需写到主LUN，因此对主LUN的性能影响较小，对于异地同步比较适用。

存储系统的异步远程复制可以支持一个主LUN有一个从LUN。

远程复制一致性组



- 用于保持多个LUN之间镜像数据的时间一致性
- 所有成员一起同步、分裂、断开和主从切换
- 每个一致性组可有8个相同复制类型的远程复制
- 主、从LUN可处于任意控制器

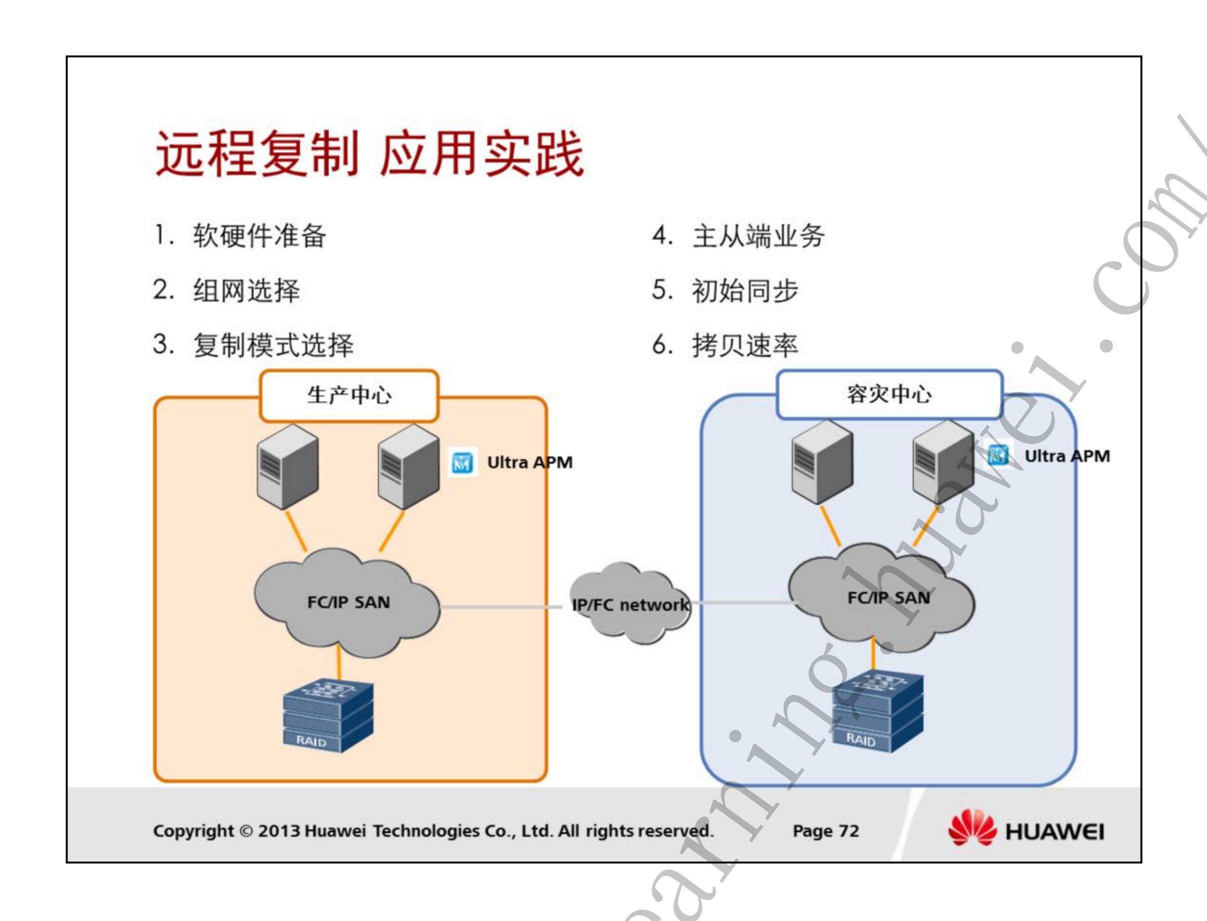
Page 71



在大中型数据库应用中，数据、日志、修改信息等存储在阵列的不同LUN中，通常称这种有关联的LUN为非独立LUN，缺少其中一个LUN的数据，都将导致其他LUN中的数据失效。

我们希望可以同时对这些LUN同时进行数据的同步或分裂等操作，以保证多个从LUN之间数据的关联性不变，从而保证容灾备份数据的完整性和可用性。

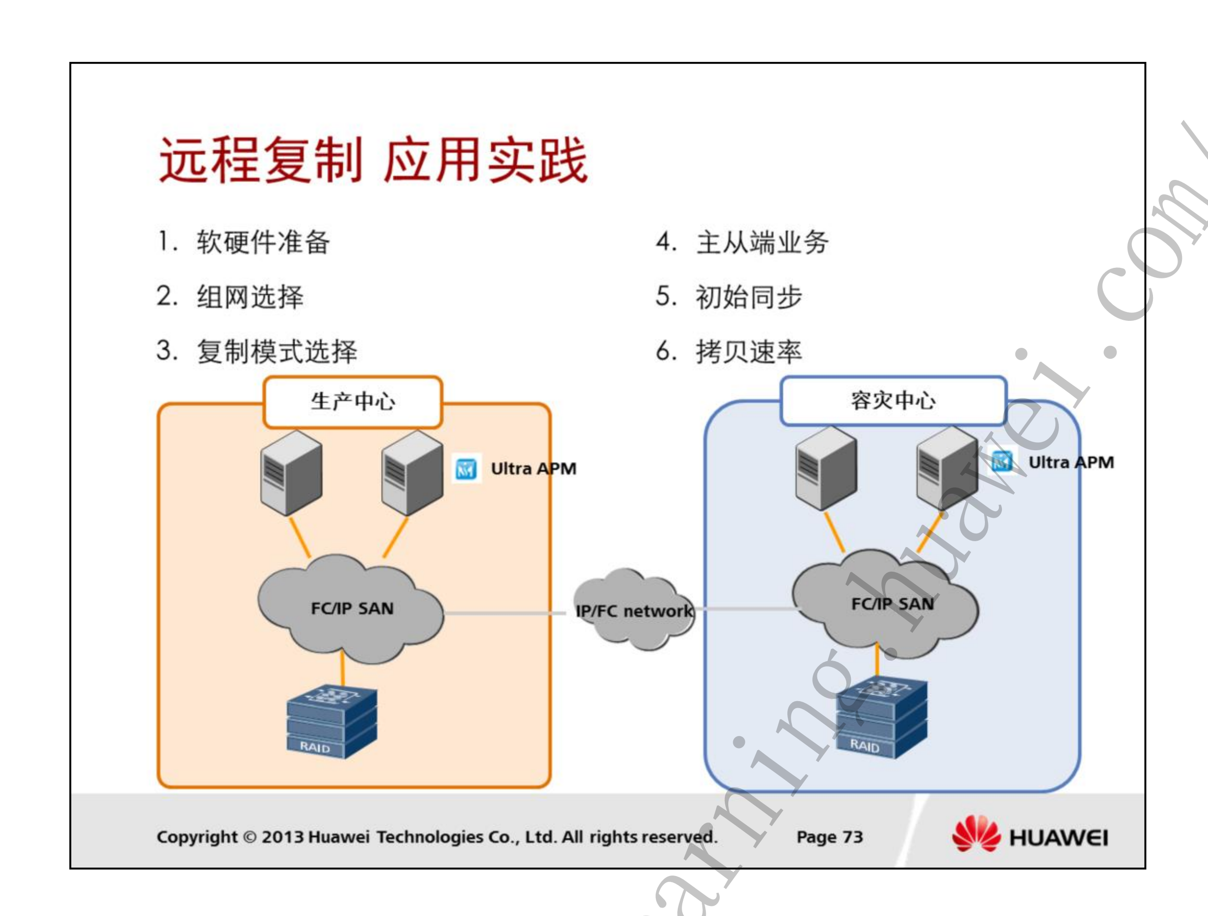
OceanStor 18000存储系统的提供的远程复制一致性组，是多个远程复制的集合。



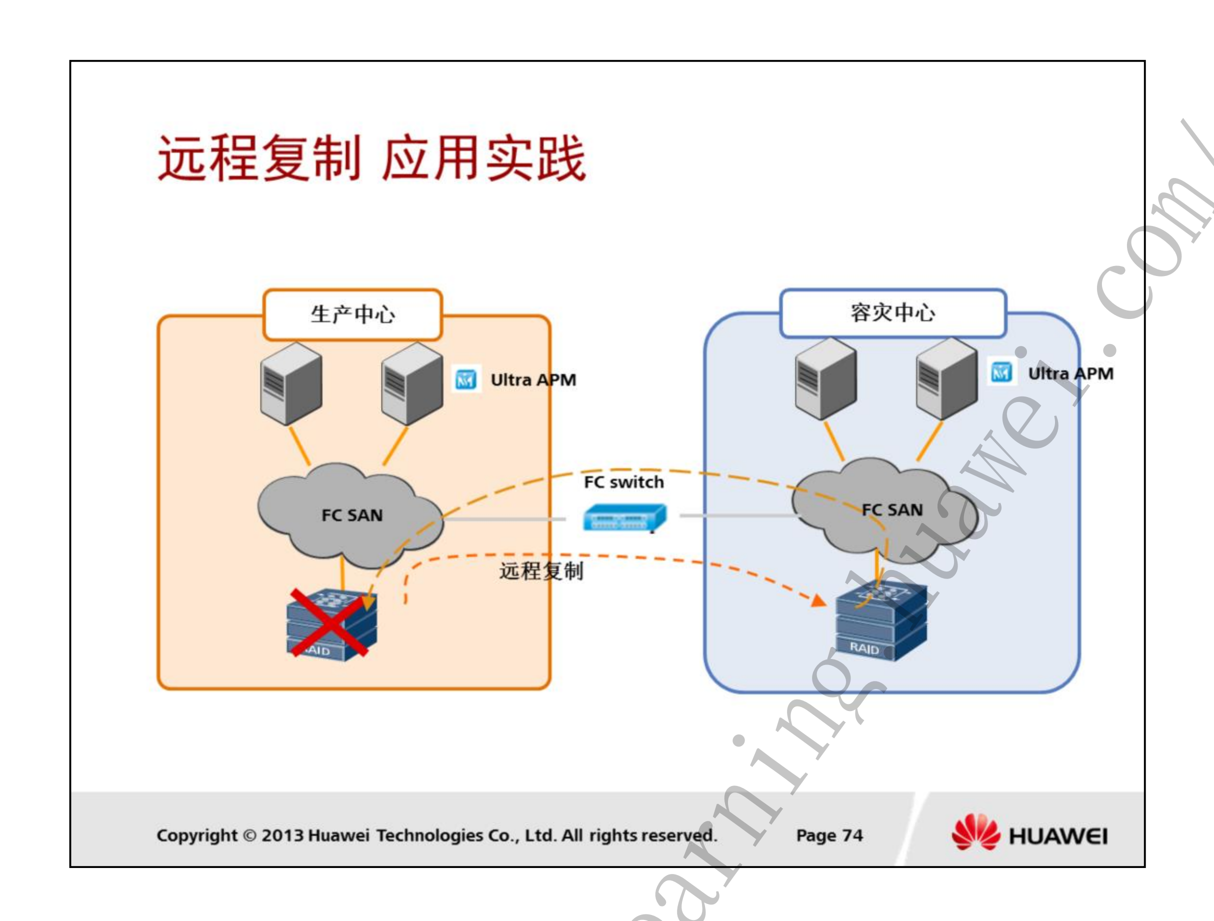
某局点生产中心的核心业务数据通过FC SAN或者IP SAN需要保存在一台OceanStor 18000设备上，同时，为了保证数据的可靠性，需要建立容灾中心，对生产中心的业务数据进行容灾备份。

容灾备份的方案包含几个步骤：

1. 软硬件准备：硬件上主要包含OceanStor 18000两台，生产中心和容灾中心服务器，FC或GE交换机；交换机用于连接生产中心和容灾中心的OceanStor 18000存储系统。服务器上除安装应用软件外，还应安装UltraAPM软件。UltraAPM安装在业务服务器上，它是与存储配套的，对存储系统和应用服务器的容灾备份特性进行统一管理的软件，为应用系统提供数据一致性保障功能。容灾中心服务器的软硬件配置要与生产中心完全一致。
2. 组网选择：如果客户环境中，容灾中心与生产中心距离较近，我们可以采用速度更快的FC组网的远程复制进行备份。 如果距离较远或者想利用现有的IP网络，可以使用iscsi组网。
3. 复制模式选择：如果用户对IO性能要求较高，一般采用异步远程复制，同步任务一般选择在业务量小的时候进行。 异步远程复制对IO性能影响较小，当灾难发生的时候会丢失部分数据，丢失数据的多少取决于上次同步开始时到当前时间的主机写IO的多少。而同步远程复制对IO性能影响较大，但是数据丢失率几乎为0，可在容灾中心与生产中心距离近且采用FC组网时采用。
4. 主从端业务：配置远程复制时，主端可以不用断业务，但从LUN不能接收主机的读写业务。



5. 初始同步：配置远程复制时，一定要选择初始同步。
6. 拷贝速率：远程复制的拷贝速率与当时的带宽和网速、所用的盘的类型、所用LUN所属的Raid组类型等都相关。



完成配置后，要使用远程复制将数据从生产中心镜像至灾备中心。生产中心发生灾难后，灾备中心接管业务。生产中心恢复后，将数据恢复到生产中心，并恢复原业务。



目录



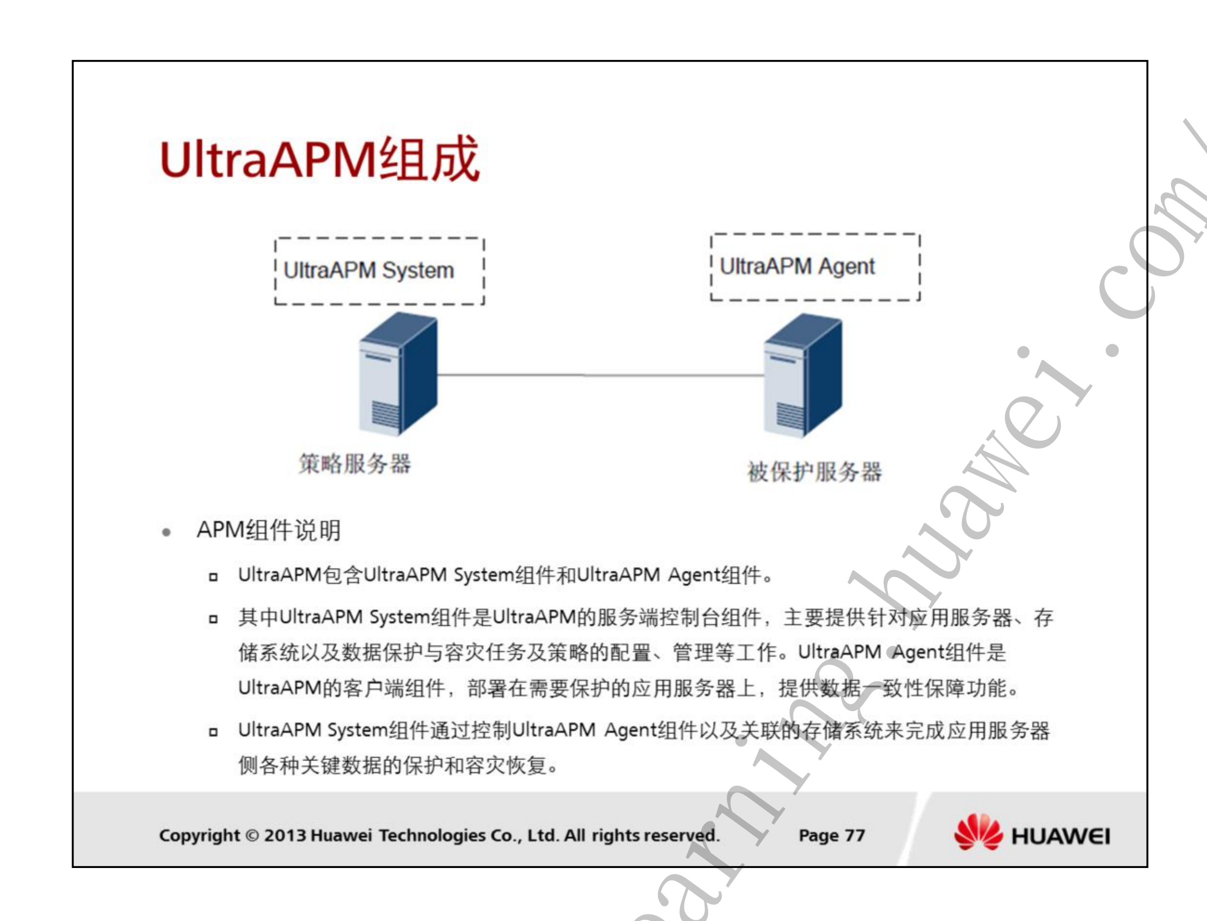
UltraAPM概述

OceanStor UltraAPM是一款华为开发的软件套件，UltraAPM利用底层存储系统所提供的增值特性，在应用服务器侧针对各类常见应用系统提供关键数据远程复制的数据保护及容灾恢复解决方案。



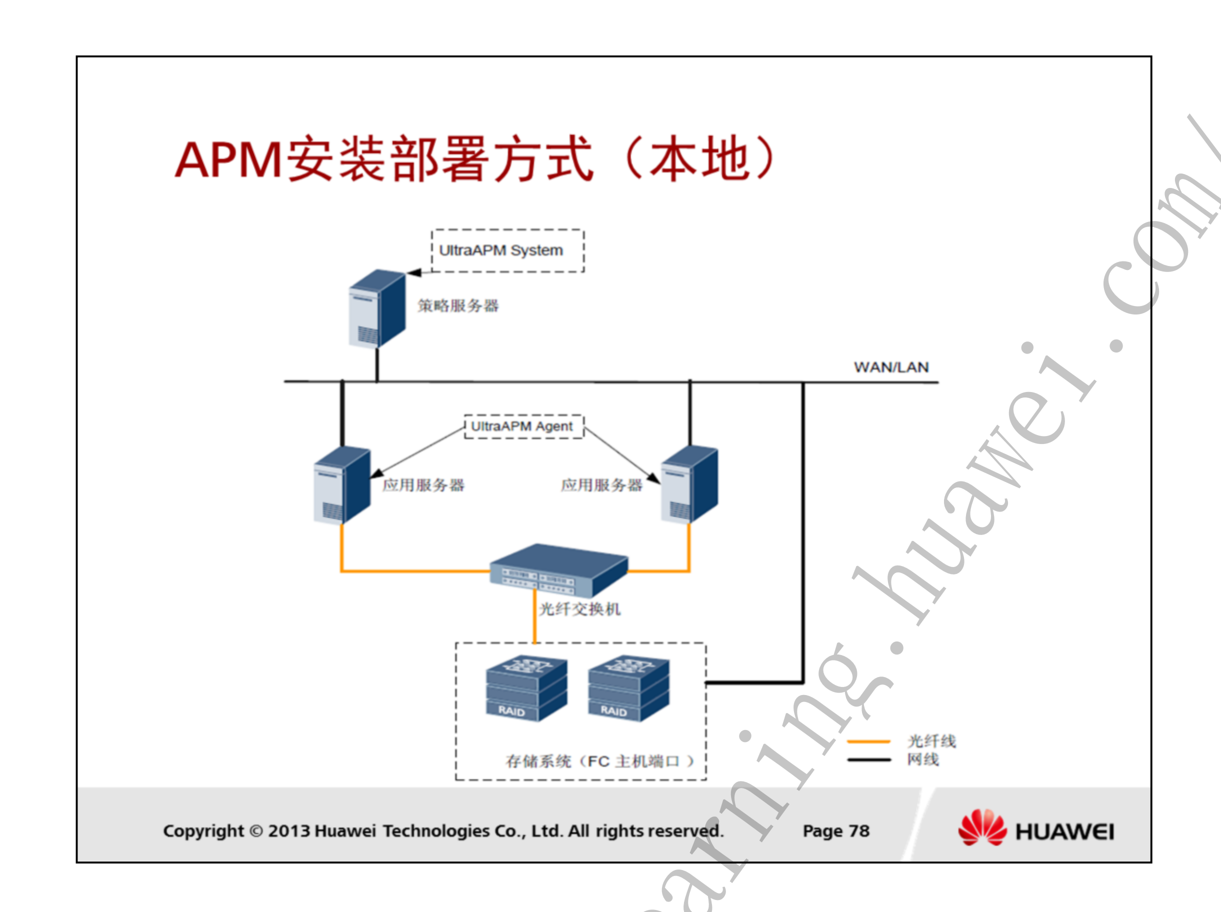
 Page 76 HUAWEI

- UltraAPM具有以下功能特性：
  - 稳妥的应用数据保护方式
  - 使用远程复制的方式对数据进行保护
  - 完善的应用程序兼容性支持Oracle、SQL Server等
  - 灵活、智能的应用系统保护策略
  - UltraAPM可自动感知服务器上已经安装的应用程序，并针对不同的数据对象定制灵活的时间策略和保护方式
  - 全面的应用系统保护方式
  - 支持任务关联功能，完全匹配真实的应用环境和应用策略
  - 方便直观的管理模式：UltraAPM提供了直观方便的管理平台“UltraAPM System”，真正实现了应用服务器、应用系统、存储系统的一站式数据保护和容灾管理



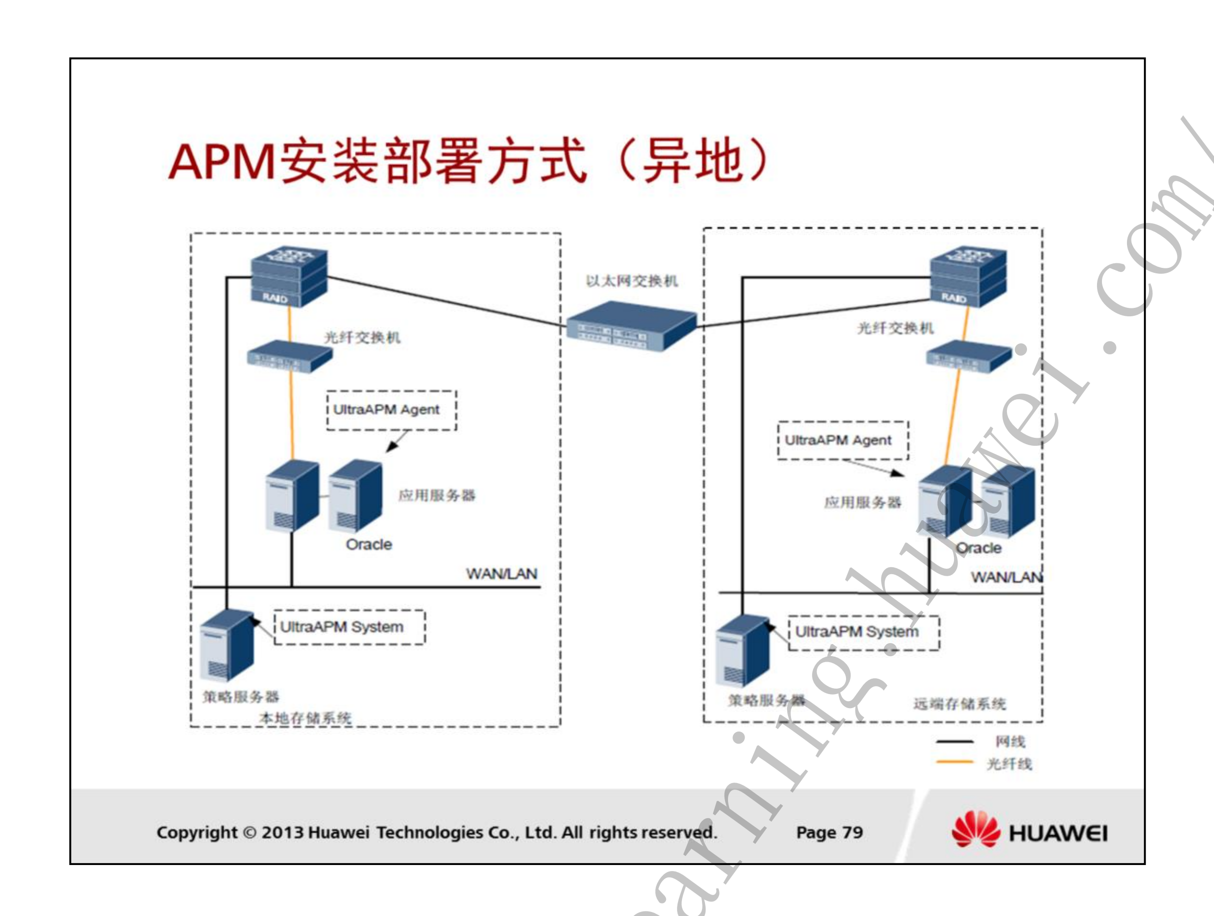
注意事项：

- UltraAPM System与UltraAPM Agent不能安装在同一台服务器上
- 所有安装了UltraAPM System组件和UltraAPM Agent组件的应用服务器均需要在APM中进行主机注册。
- 不能随便修改APM主机的主机名，如果修改了主机名，又改了主机的IP，会导致软件不能用。



本地容灾部署说明：

- 支持包括集群环境和非集群环境。
- UltraAPM Agent组件必须正确安装在各个应用服务器上（每台服务器）。
- 策略服务器、应用服务器和存储系统必须确保能互相连通（通过IP网络）。
- 集群环境支持类型:A/A集群组、A/P集群组以及Oracle RAC组



- 客户设备IP地址为私网地址，隐藏在内部网络中。
- 支持包括集群环境和非集群环境。
- UltraAPM Agent组件必须正确安装在各个应用服务器上。
- 策略服务器、应用服务器和存储系统必须确保能互相连通；本地存储系统与远端
- 存储系统必须确保互相连通。

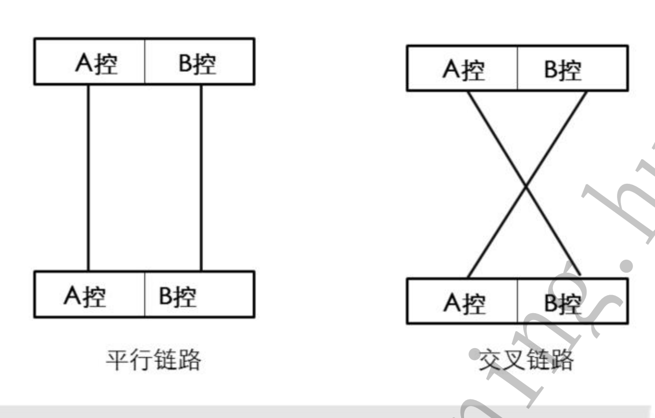


目录



链路概念

- 链路指阵列直接进行数据传输的通道，一般都是阵列主机接口直接的连接
- 链路按照线路可以分为IP链路和FC链路
- 链路按照控制器的对应关系分为平行链路和交叉链路

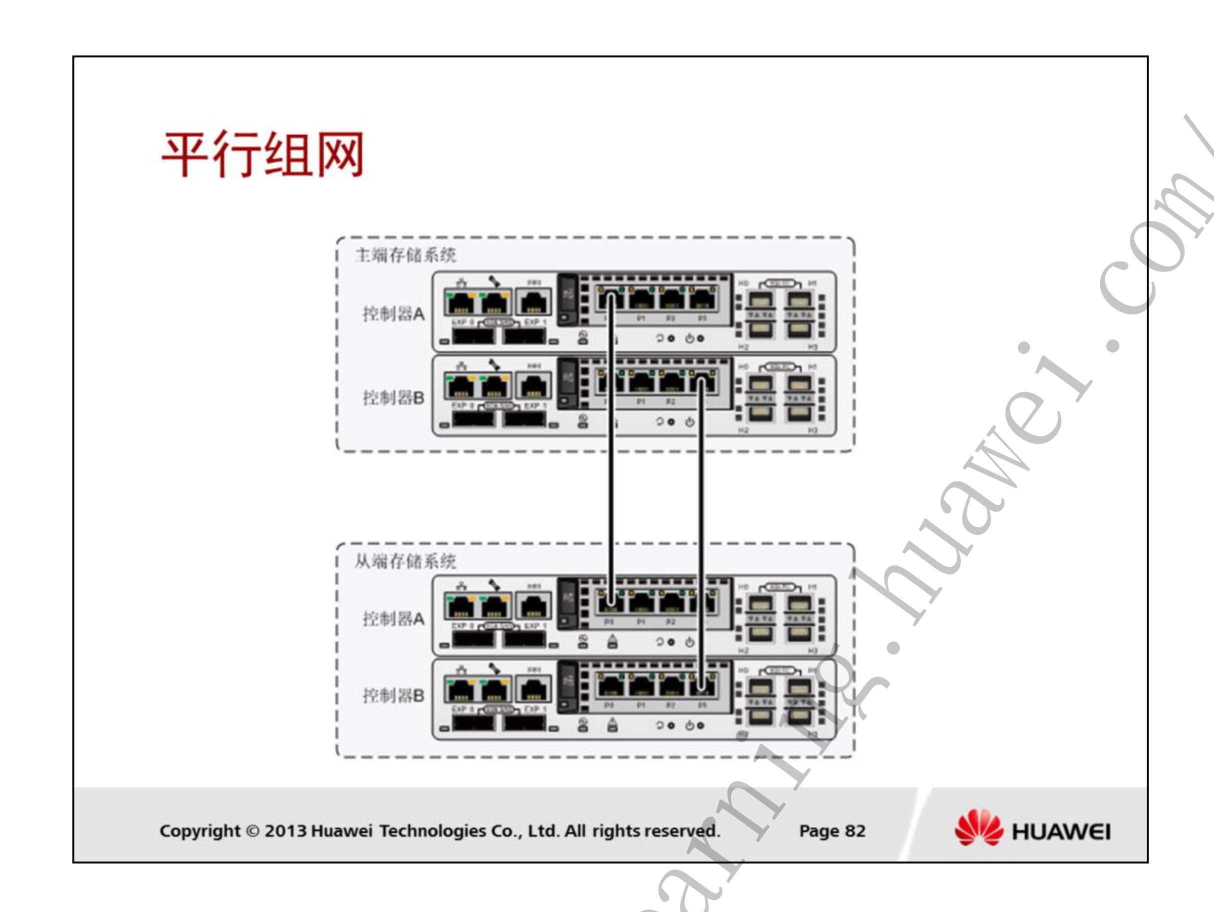


Page 81

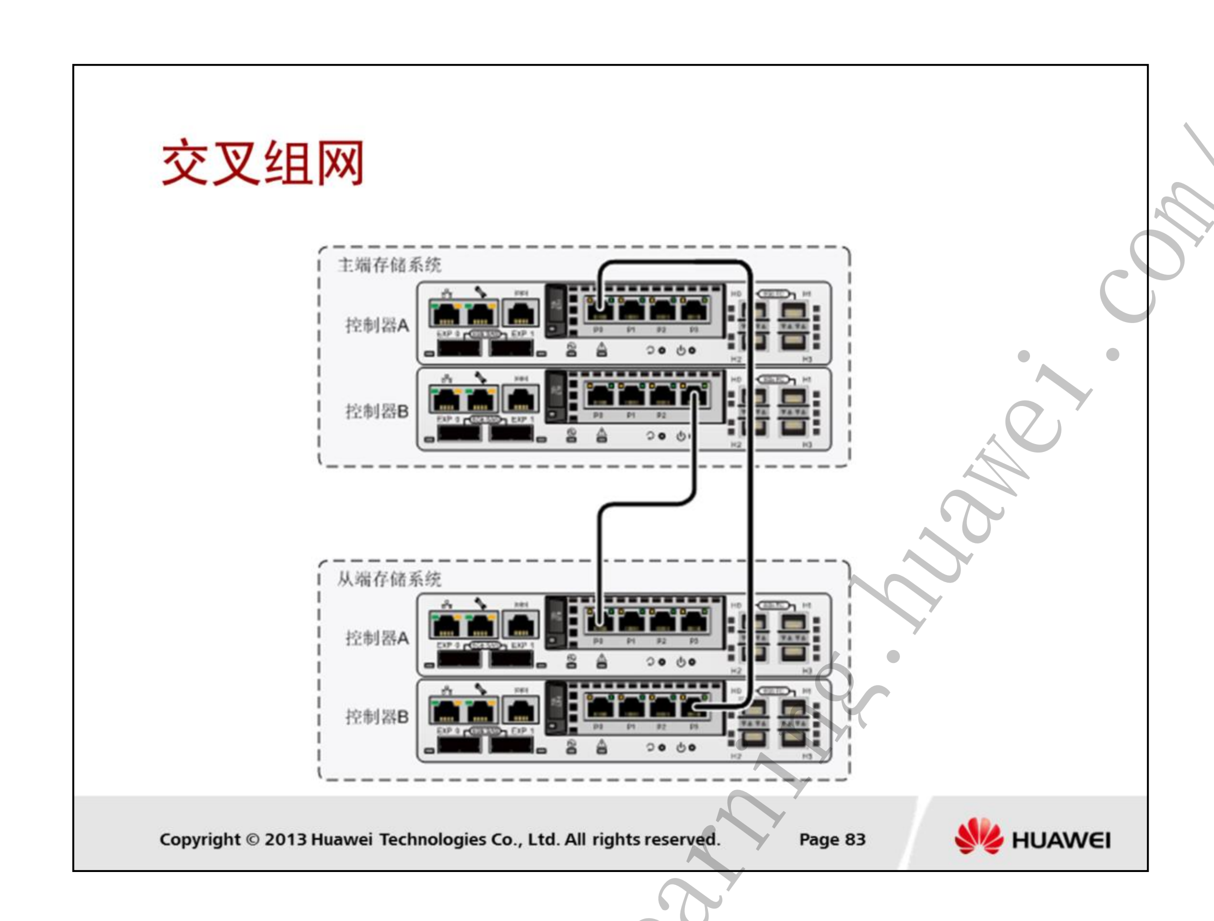


在进行组网时，两台存储设备之间不能同时使用FC和iSCSI进行组网。

为了提高系统的性能，建议将所有主LUN或从LUN创建在一个控制器下，然后选择平行组网或交叉组网的方式，使拥有所有主LUN的控制器与拥有所有从LUN的控制器连接在一起。



主端存储系统控制器A的主机端口与从端存储系统控制器A的主机端口连接，或者主端存储系统控制器B的主机端口与从端存储系统控制器B的主机端口连接。平行组网的示意图下所示（以S5500T存储系统的iSCSI组网为例）。



主端存储系统控制器A的主机端口与从端存储系统控制器B的主机端口连接，或者主端存储系统控制器B的主机端口与从端存储系统控制器A的主机端口连接。交叉组网示意图如下所示（以S5500T存储系统的iSCSI组网为例）。



- 本端网口ID： 选择用于连接对端的本端iSCSI口的ID号，可以在“端口”界面中查询到
- 目标器IP地址：输入对端的iSCSI口的IP地址。
- 其余配置项根据实际情况进行修改。

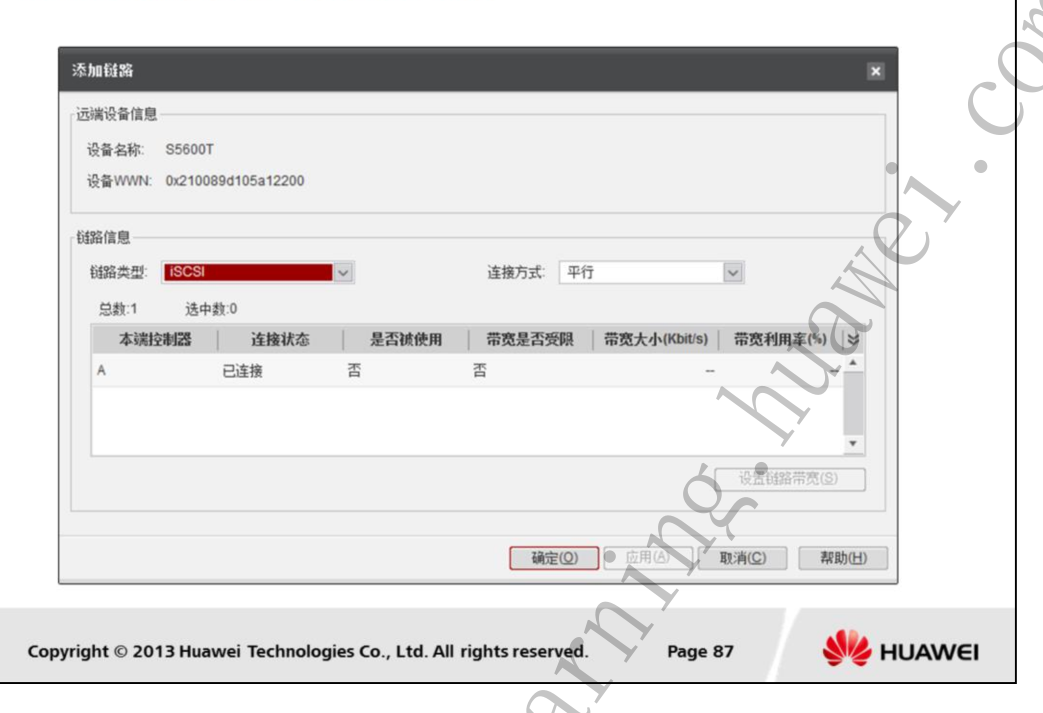


如果两台阵列间通过IP链路进行连接，需要在两台阵列的端口设置中添加对方为目标器。

IP链路管理-远程连接管理

 Page 86 HUAWEI

添加目标器之后，可以在目标器管理中查看到已经添加的目标器信息，此时的状态时“未连接”，需要点击“连接”，同时，需要在两台阵列上分别确认链路状态是否已连接。



- 在链路类型选择iSCSI
- 在连接方式选择平行或交叉



如果两台阵列的FC主机口之间通过光纤网络已经连通，并且端口协商已经成功，那么两台之间的FC链路就已经连接。



如果两台阵列间的链路已经建立成功，那么需要进行远程连接管理，指定两台阵列间需要通过哪条链路进行连接。

两台阵列间的要么只能通过FC链路进行连接，要么只能通过iSCSI链路进行连接。



添加链路完成后，阵列之间的链路管理完成。



思考题

1. RAID2.0同传统RAID的区别是什么？
2. 华为存储有哪些功能用来提升客户体验？





总结

- 统一存储产品形态及功能特性
- 主机同存储联动管理
- 链路管理和组网





习题

- 判断题
  1. SmartTier每个存储层里可以有不同硬盘类型（T of F)
- 多选题
  1. 以下哪些是RAID2.0的数据组织组件？（　　）

     A. chunk

     B. CKG

     C. extent

     D. LUN

 Page 93



- 习题答案：
  - 判断题：1.F
  - 单选题：1.ABCD

Thank you

www.huawei.com

HC120920003 虚拟化存储网关系统部署与管理




HC120920003
存储虚拟化技术及应用
www.huawei.com
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



目标

- 学完本课程后，您将能够：
  - 熟悉存储虚拟化的概念及其价值
  - 熟悉存储虚拟化技术实现方式及其优缺点
  - 了解虚拟化存储网关系统架构及原理
  - 掌握虚拟化存储网关系统的基本业务配置和操作
  - 了解虚拟化存储网关系统高级技术

 Page 1





目录



存储虚拟化定义

什么是存储虚拟化? —SNIA 根据以下几个角度进行定义

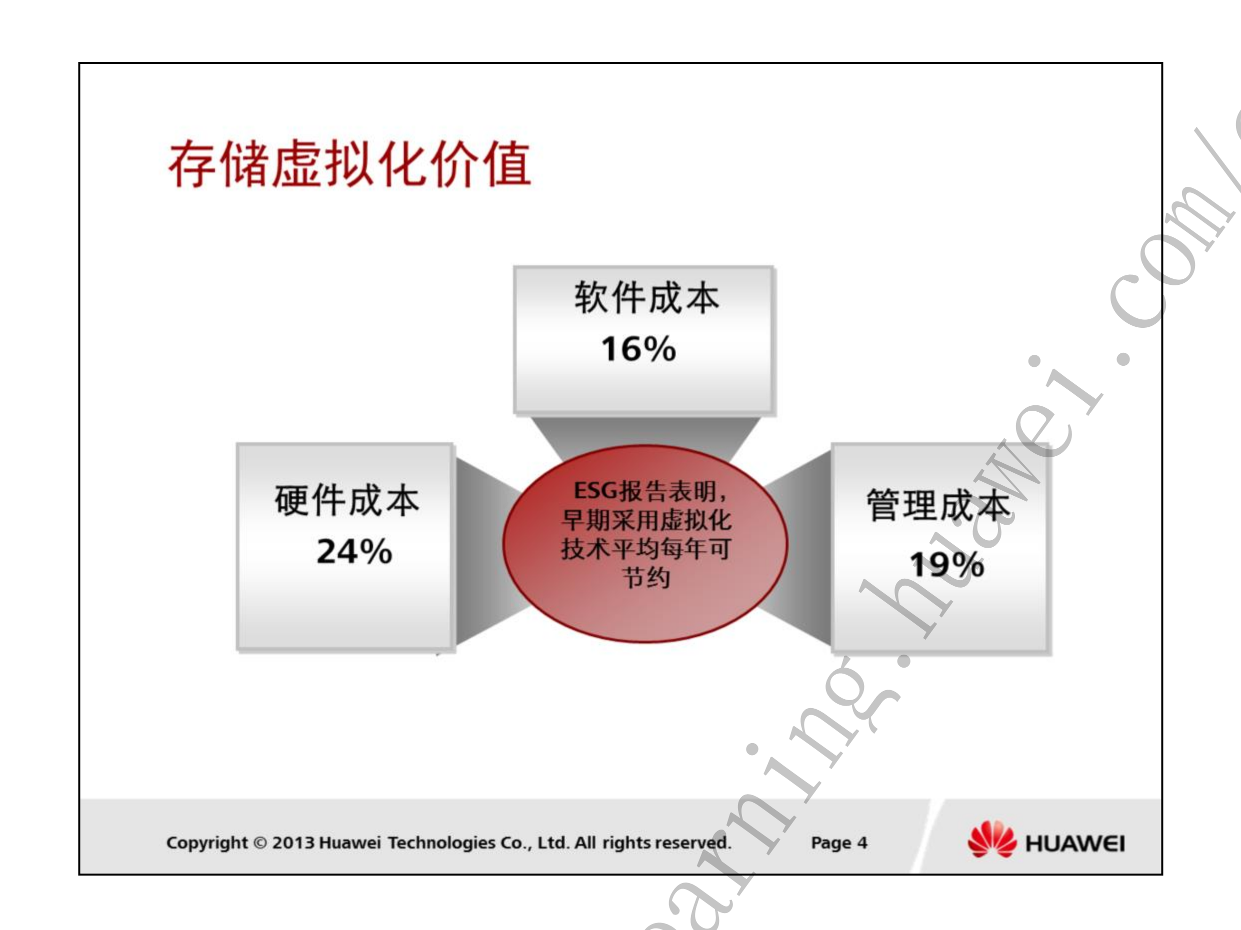
- 在不同的实现层次:
  - 主机层实现(Application, OS,HBA)
  - 网路层实现(Switch, Router, Gateway)
  - 存储层实现(Array , Library Device)
- 对主机透明化:
  - 物理路径
  - 设备规格
  - 数据存放物理位置
- 实现数据存放位置及部署的透明化
- 动态操作使能:
  - 实现不中断业务重配置
  - 数据存放位置对主机环境透明
- 存储虚拟化是分层，分级和多样化的

 Page 3



虚拟存储技术将主机应用层、网络层和存储层设备进行抽象化统一管理，向服务器层屏蔽存储设备的特殊性，而只保留其统一的逻辑特性，从而实现了存储系统集中、统一而又方便的管理。存储虚拟化由多种不同的类型、方法来实现并且虚拟化程度不一。对比一个计算机系统来说，整个存储系统中的虚拟存储部分就像计算机系统中的操作系统，对下层管理着各种特殊而具体的设备，而对上层则提供相对统一的运行环境和资源使用方式。

虚拟化是一种实现对逻辑环境进行简单管理的有效手段。通过虚拟化，用户将摆脱底层物理环境的复杂性，充分利用基于异构平台的存储空间，在开放的基础上实现对资源的有效规划。虚拟化可以自动配置存储设备及其空间，使用户能在一个域中使用在物理上分散存在的所有存储资源，以便跨地区管理不可预测的事件，如业务不连续性、对容量需求的调整、员工的变化等。而无论这些存储资源所处的存储域的位置、大小、类型和制造商如何，都将被从单一逻辑视图中进行管理。

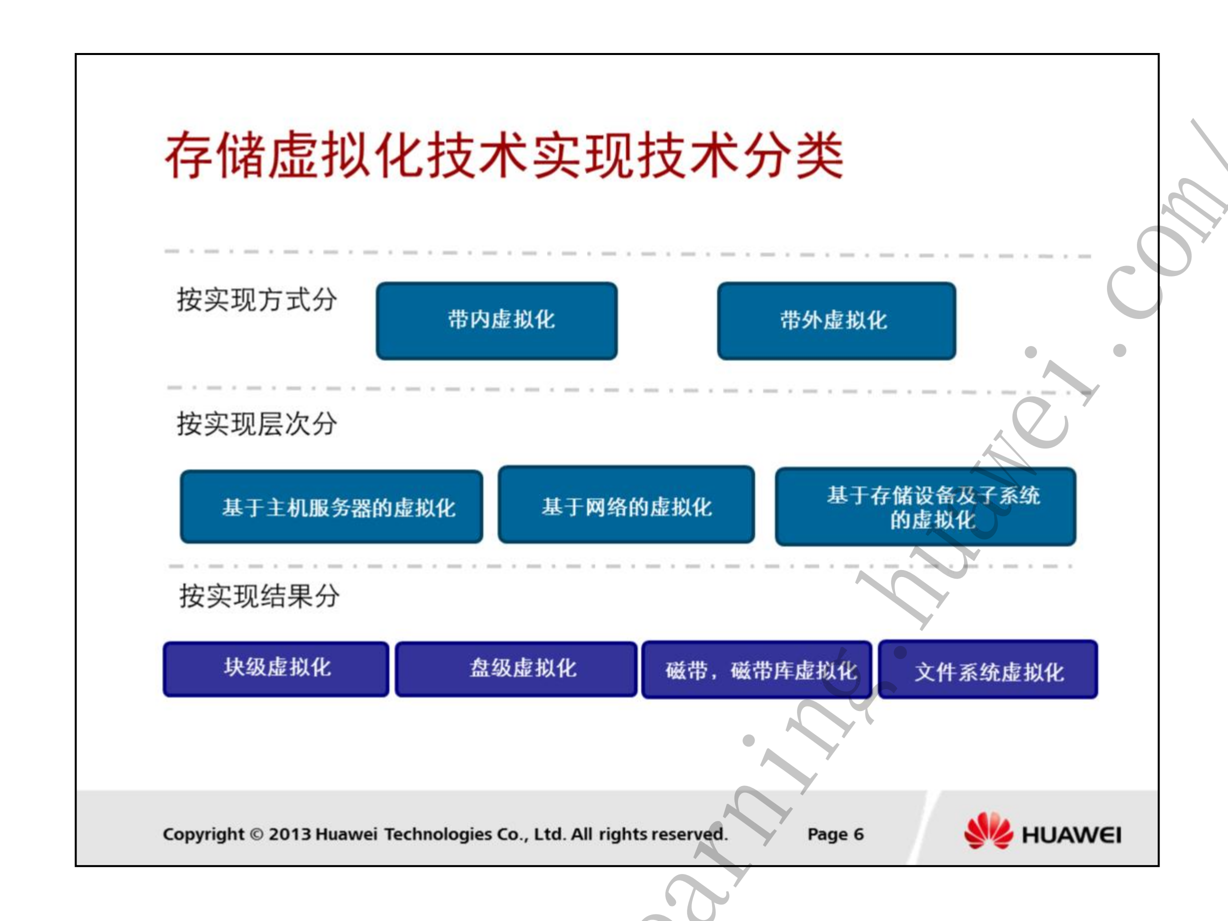


- 成本节约、高效、简化。
    - 在传统的开放系统环境中，主机系统拥有自己的存储资源，每个系统管理员需要针对每一种系统来控制存储资源的分配、使用和管理。此外，每一台服务器还拥有一 个单独配置和管理的文件系统。据有关权威调查显示，传统的存储环境中，开放系 统的容量利用率仅为40%至50%，这种低使用率造成了企业不得不耗费大量的额外成本来满足存储扩容的需求。
    - 因此随着存储技术的不断发展，存储虚拟化技术与方案以其真正意义上的开放，在众多IT技术中脱颖而出。在虚拟存储环境下，存储对用户来说将变得透明，用户可以不必关心存储设备的功能差别、容量大小、设备类型和制造商如何，所有的设备将被统一管理，而且赋予统一的功能如Flashcopy、远程灾备等。今天，存储虚拟化已经不再是一个陌生的词汇，越来越多的企业开始关注存储虚拟化，同时将其实施到IT规划中。



目录





- 实现方式：带内虚拟、带外虚拟。
- 实现层：基于主机的虚拟、基于网络的虚拟化 基于存储设备、存储子系统的虚拟化。
- 实现结果：块虚拟、磁盘虚拟、 磁带、磁带驱动器、磁带库虚拟、文件系统 虚拟化。

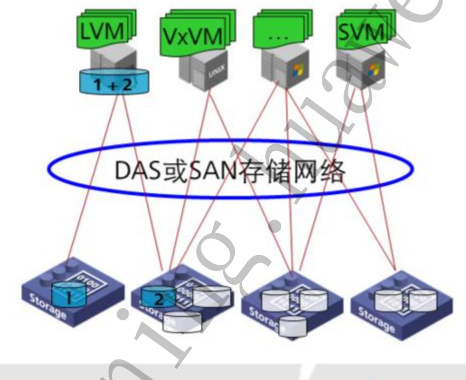


目录



基于主机层的存储虚拟化

- 应用场景：多个异构的阵列的LUN在服务器上被整合为一个大的LUN使用，常用在不同磁盘阵列间做数据镜像保护，容量扩容等。
- 实现方式：一般使用操作系统的逻辑卷管理软件，不同操作系统的逻辑卷管理软件也不相同。
- 常见产品： LVM， VxVM



 Page 8



不同系统有自己的卷管理软件。像Solaris使用SVM，HPUX和AIX使用LVM卷管理软件。

卷管理方式还可以通过第三方软件进行管理，如Symantec的VxVM软件可以在不同的系统平台上进行磁盘卷管理。

卷管理软件可以使服务器的存储空间可以跨越多个异构的磁盘阵列，比如LUN 1 和LUN 2 来自不同的阵列，在主机上由LVM卷管理软件合成一个大LUN（LUN 1 + LUN2）。同时也可以通过卷管理软件提供丰富的数据管理能力。

- 优点： 支持异构的存储阵列。
- 缺点：
  - 占用主机资源多，降低应用性能。
  - 主机升级、维护和扩展非常复杂， 且容易造成系统不稳定性。
  - 需要复杂的数据迁移过程，影响
  - 业务连续性。

LVM概念及架构

LVM，逻辑卷管理器 (Logical Volume Manager)，它是用于管理逻辑卷的，LVM 在 Linux 内核得到支持。

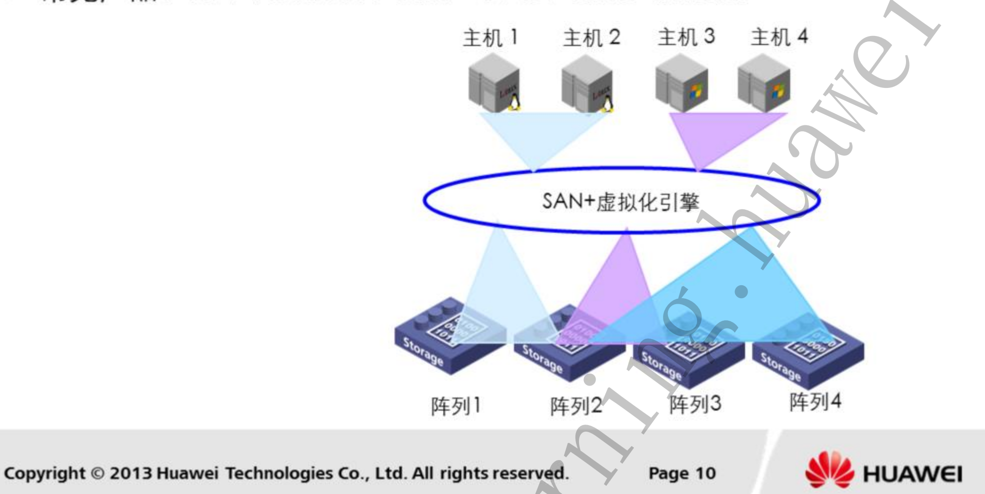
文件系统	EXT3	XFS	Reiserfs	未分配
LVM层	LV	LV	LV	
	LE LE LE	LE LE LE LE LE	LE LE LE LE	
	Volume Group			
	PE PE PE PE	PE PE PE PE	PE PE PE PE	PE PE PE PE
	PV	PV	PV	PV
物理分区	PP	PP	PP	PP

Page 9



在传统的存储模型中，文件系统是直接构建于物理分区之上的，物理分区的大小就决定了其上文件系统的存储容量，因此对文件系统的存储容量的调整就变得比较繁琐。而 LVM 设计的主要目标就是实现文件系统存储容量的可扩展性，使对容量的调整更为简易。

PP，物理分区（Physical Partition），如硬盘的分区，或 RAID 分区。PV，物理卷（Physical Volume），是 PP 的 LVM 抽象，它维护了 PP 的结构信息，是组成 VG 的基本逻辑单元，一般一个 PV 对应一个 PP。PE，物理扩展单元（Physical Extends），每个 PV 都会以 PE 为基本单元划分。VG，卷组（Volume Group），即 LVM 卷组，它可由一个或数个 PV 组成，相当于 LVM 的存储池。LE，逻辑扩展单元（Logical Extends），组成 LV 的基本单元，一个 LE 对应一个 PE。LV，逻辑卷（Logical Volume），它建立在 VG 之上，文件系统之下，由若干个 LE 组成。



阵列1 阵列2 阵列3 阵列4被虚拟化引擎统一管理，资源通过虚拟化引擎可任意映射不同的主机。

- 优点：
  - 与主机无关，不占用主机资源。
  - 能够支持异构主机、异构存储设备。
  - 统一管理不同存储设备的数据。
  - 构建统一管理平台，可扩展性好。
- 缺点：非主流厂商产品成熟度较低，存在和不同 存储和主机的兼容性问题。

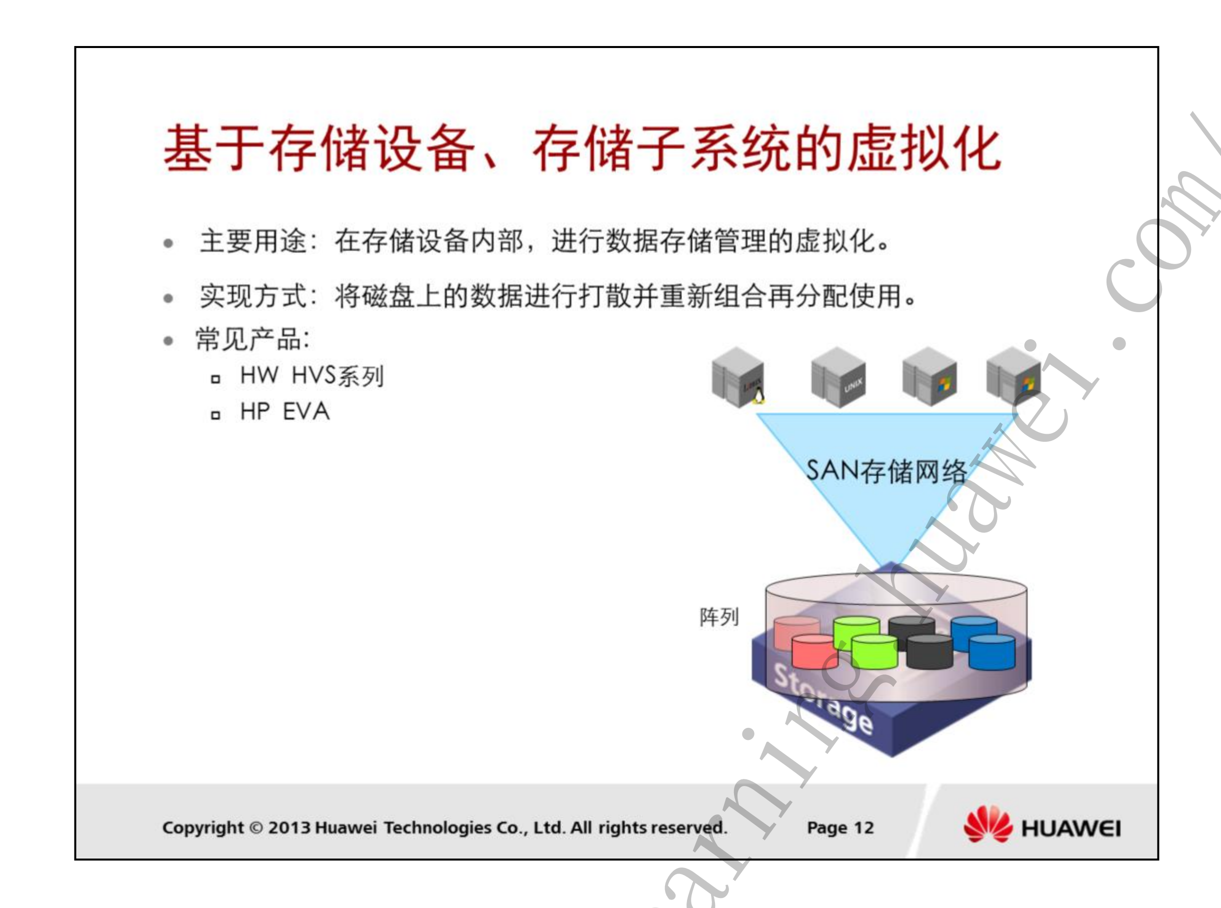
华为 VIS6600T

- 虚拟化能力
- 异构阵列资源整合
  - 网络层虚拟化架构，能将不同品牌、不同型号、不同架构的存储设备整合为统一的存储资源池，用户只需要关注 VIS6600T提供的逻辑存储资源，而不必关心后端存储的具体形态
- FC和IP的无缝融合
  - 前端支持FC和iSCSI主机端口，后端支持FC SAN阵列和IP SAN阵列
- 资源按需分配
  - 允许应用程序向用户提供的容量多于在存储系统中实际分配的容量，提高了存储容量利用率，使资源调配更简化，容量分配更灵活，可降低总体拥有成本

 Page 11

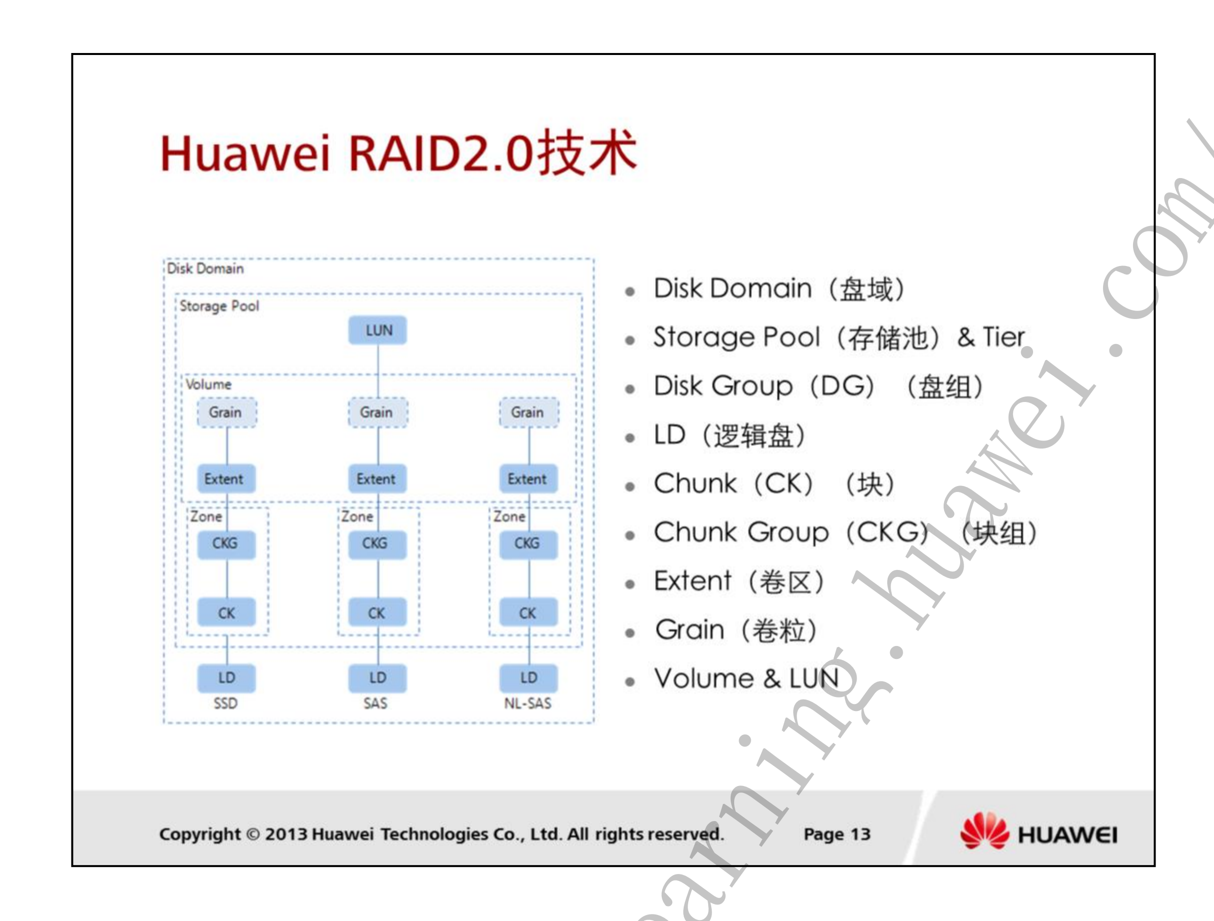


- VIS6600T是一款基于网络层实现的存储虚拟化产品。
- 它的主要优势在于：
  - 兼容异构的主机操作系统和存储磁盘阵列。
  - 基于日志机制的复制技术，保障容灾数据的一致性。
  - 无需在主机端安装容灾软件，同时不影响主机性能。
  - 可提供基于存储虚拟化的数据迁移方案，以解决数据首次同步占用较多带宽和时间的问题。
  - 可扩展性强，新增主机或存储设备可快速便捷地接入现有容灾网络。



阵列1可以将阵列2和阵列3到阵列N的资源整合，对主机统一由阵列1对外提供资源。

- 优点：
  - 与主机无关，不占用主机资源和网络资源。
  - 数据管理功能丰富。
- 缺点：
  - 对阵列处理性能要求更高。



从虚拟化的角度来看，RAID2.0技术同之前传统的RAID技术有很大的不同，传统的RAID对于数据存储的管理也是虚拟化，但是其最小管理单位是整块硬盘。而RAID2.0是将单块硬盘进行了打散，将数据存储管理的粒度大大减小，使得数据的组织和管理更加精细化。

对存储系统这一底层部件的虚拟化技术革新大大提升了数据保存的可靠性和管理的灵活性。

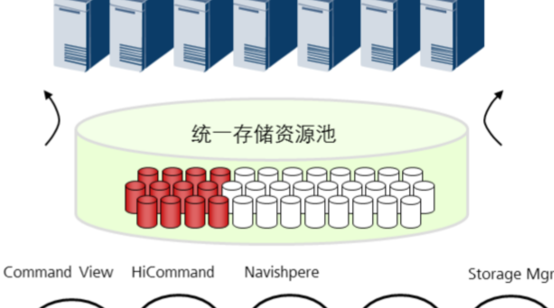


目录



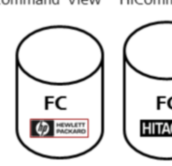


- 将不同协议（iSCSI、FC）、不同架构（IP SAN、FC SAN）、不同品牌的存储设备整合到统一的存储资源池



存储资源物理位置、主机对存储阵列的访问路径、设备物理特性等对主机不可见

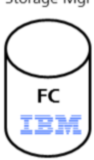
Command View　HiCommand　Navishpere　Storage Mgr







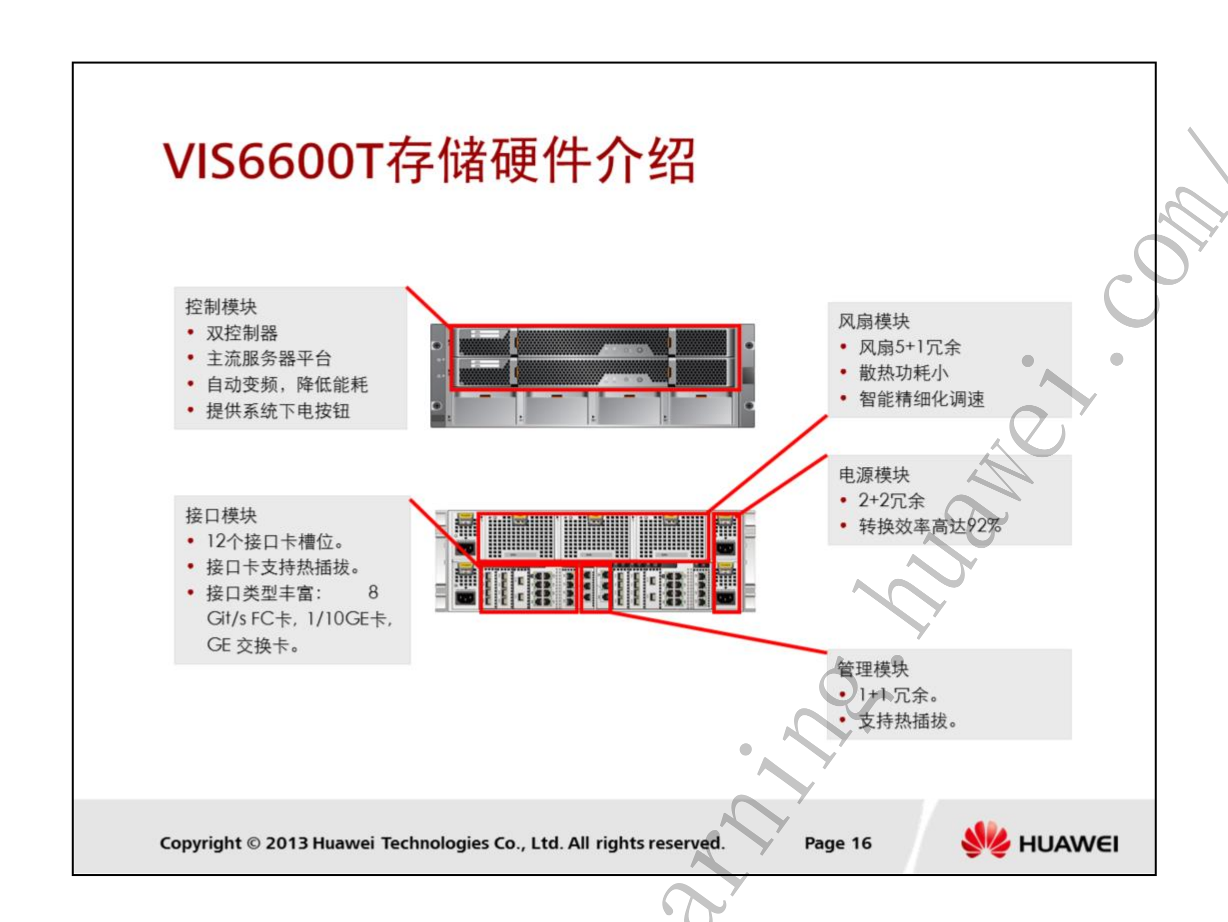




　Page 15



- 存储资源物理位置和具体现实透明化，访问路径、设备物理特性等对主机不可见。
- 允许物理存储架构动态变化。
- 从存储资源池中动态分配存储资源，并实现存储集中管理。



先进的设计：模块化设计保证硬件灵活性.

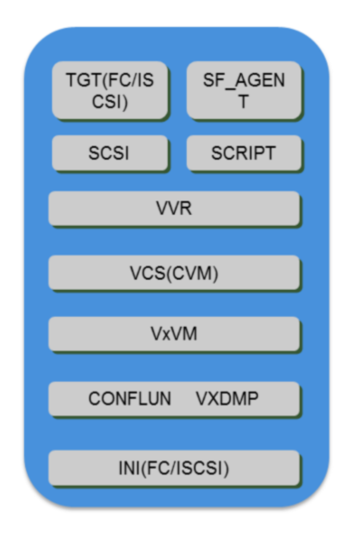
集群架构、全冗余、全模块化设计、实时监控告警，多种业务接口满足不同业务需求，保证硬件灵活性、可靠性、扩展性。

VIS6600T产品规格

	VIS6600T
处理器	XEON 5645
内存容量	48GB
前端端口类型	8Gb FC和1/10GE(iSCSI)
后端端口类型	8Gb FC和1/10GE(iSCSI)
最大业务端口数	20(8GbFC) / 10(10GE) / 20(GE)
最大卷数目	4096
最大LUN数目	1024
最大主机数量	1024 FC / 256 iSCSI
最大卷容量	64TB
节点数量	2 / 4 / 8
功耗	480W



软件介绍



模块名称	模块概述
SF_AGENT	提供给用户管理子系统和SF之间的通信
SCRIPT	对SF软件所有功能的脚本封装,提供自研模块和第三方模块的交互;
TGT(FC/ISCSI)	支持FC/ISCSI协议的目标器驱动,提供VIS和主机的通信通道,提供主机管理和卷映射功能,响应用户查询设置操作
SCSI	提供SCSI协议的支持,配合实现IO多路径功能
VCS(CVM)	提供在VIS控制器间的集群功能
VVR	提供基于IP的远程复制功能
VxVM	SF软件功能的基础,提供增值业务
CONFLUN	保存磁盘组和卷的信息
VXDMP	负载均衡、路径选择、分流配置IO和业务IO
INI(FC/ISCSI)	支持FC/ISCSI协议的启动器驱动,提供VIS和阵列的通信通道,识别阵列提供的LUN
DB	对INI/SCSI/VXDMP等模块提供配置访问、修改和保存功能



VIS6600T存储虚拟化

- VIS6600T为数据管理系统提供了一层虚拟的“卷”的逻辑设备，来屏蔽异构存储设备的差异。

- VIS的卷管理软件采用的是VxVM 。
- VIS6600T使用两种对象进行存储管理：物理对象和虚拟对象。
  - 物理对象是指物理磁盘，或者类似磁盘的存储设备，是数据最终的存放地。
  - 虚拟对象是VIS6600T进行存储设备管理的逻辑对象。VIS6000通过虚拟对象和物理设备的映射来访问存储设备。
- VIS提供的虚拟化技术可以保证：
  - 不对存储在原设备LUN上的数据内容做任何修改。
  - 不对存储在原设备LUN上的数据位置做任何修改。
  - 可以在用户现有环境中无缝接入VIS。
  - 原LUN可以随时恢复和原主机的配置关系。
  - 对上层的应用无任何影响。

VIS6600T软件架构

- VIS6600T集群卷管理系统CVM
  - 用于管理存储设备的系统
  - 提供A/A的访问模式
- VIS6600T存储管理软件ISM
  - 为VIS6600T的用户提供GUI管理方式

HOST DISK

DISK GROUP

VOLUME

PLEX

VM DISK

VM DISK

LUN

LUN

VIS6600T内部模块



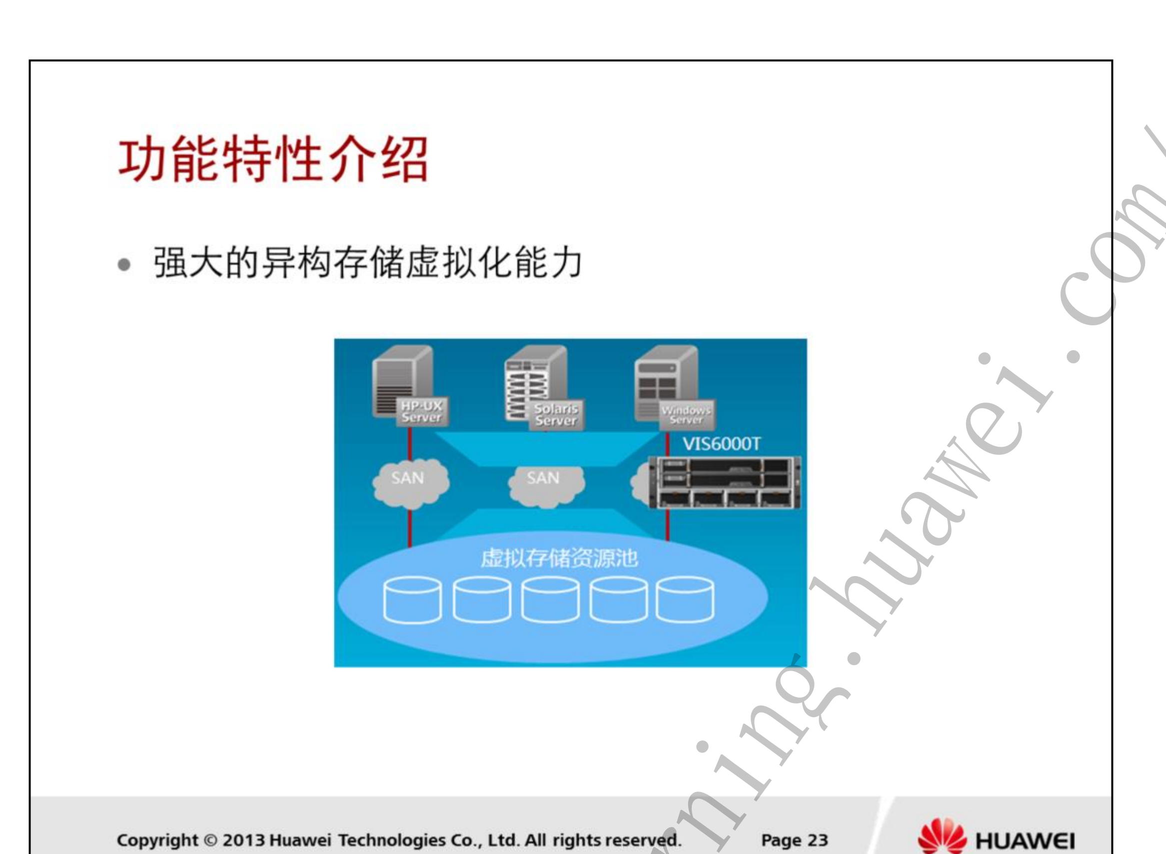
目录



功能特性介绍

- 强大的异构存储虚拟化能力
- 良好的可靠性和运行稳定性
- 支持Scale-out扩展，性能线性增长
- 灵活的业务与数据保护功能





- 业界领先的广泛兼容性，能够兼容业界主流的存储设备；
- 基于网络层的异构存储虚拟化技术，能够将不同厂商存储整合为统一的存储资源池，实现存储资源共享和统一管理；
- 原有存储数据无需迁移和转换。

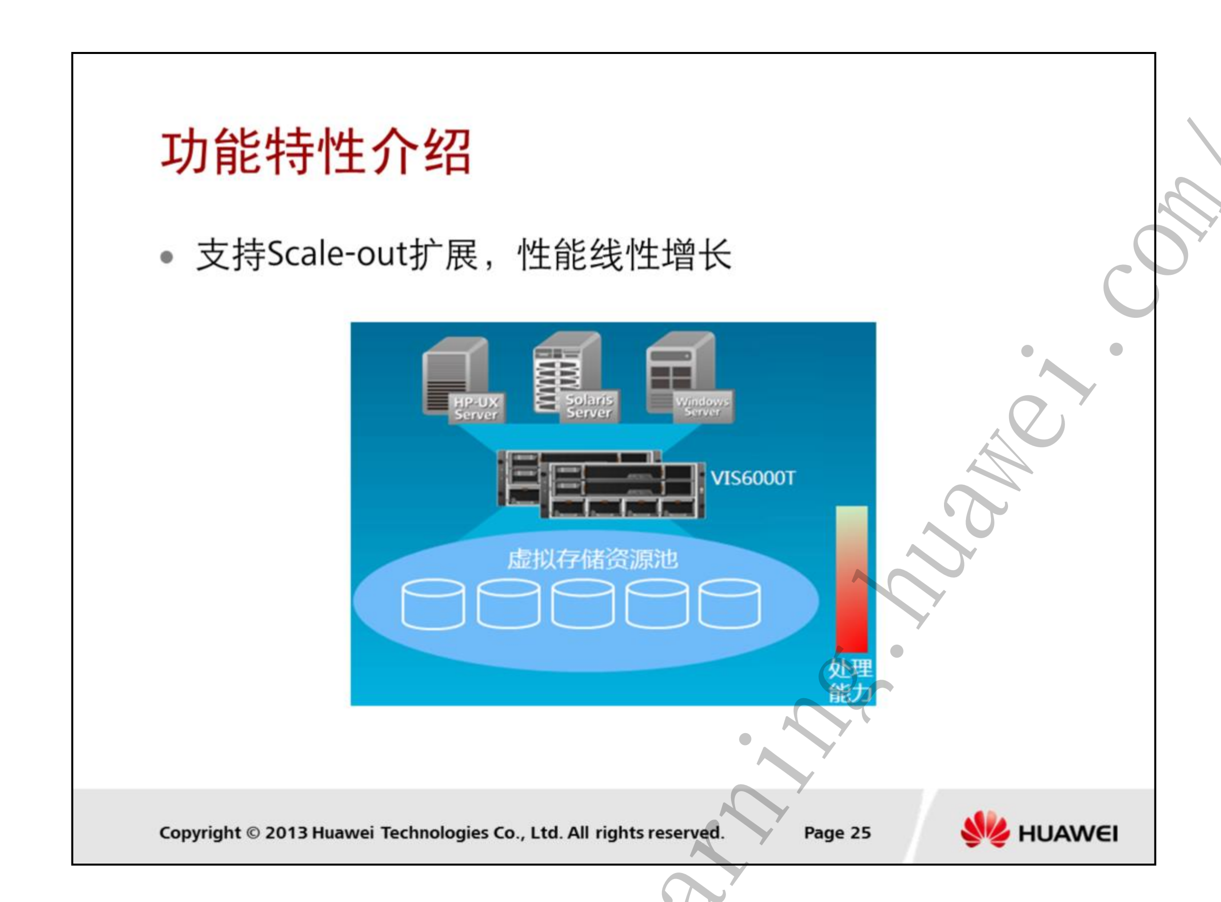
功能特性介绍

- 良好的可靠性和运行稳定性

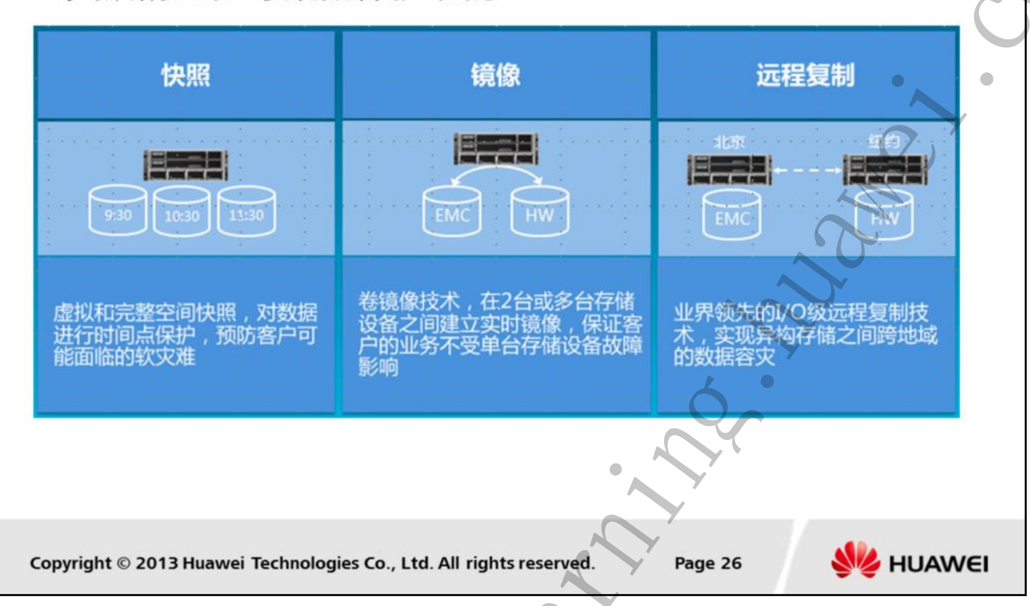


Page 24





- 业界领先的Scale-out扩展方式，通过扩展集群内节点数量线性提升整体的业务处理能力；
- 客户可以按需购买节点，降低了初始购置成本压力。


功能特性介绍
灵活的业务与数据保护功能
快照
镜像
远程复制
9:30
10:30
11:30
EMC
HW
北京
纽约
EMC
HW
虚拟和完整空间快照，对数据进行时间点保护，预防客户可能面临的软灾难
卷镜像技术，在2台或多台存储设备之间建立实时镜像，保证客户的业务不受单台存储设备故障影响
业界领先的I/O级远程复制技术，实现异构存储之间跨地域的数据容灾
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 26
HUAWEI



目录



VIS6600T基本配置—命名规则

磁盘别名命名规则

根据用户阵列原有LUN名字来命尽，

中间加磁盘阵列编号，以及Lun ID编号，最后以DSK结尾。

卷命名规则

根据磁盘名字来命名，以vol结尾。

磁盘组命名规则

根据应用类型来命令，以dg结尾。

RVG命名规则

根据应用类型来命令，以RVG结尾。

SRL命名规则

根据应用类型来命令，以SRL结尾。

DCM磁盘命名规则

根据应用类型来命令.以DSK结尾。

DCO磁盘命名规则。

根据应用类型来命令.以DSK结尾。



磁盘别名命名规则

例如：Data_S01_L01_DSK1、

Lock_S01_L02_DSK1、

Quorum_S01_L02_DSK

卷命名规则

例如：Data_S1_L1_VOL、

Lock_S1_L2_VOL、

例如：Oralce_DG、Exchange_DG

RVG命名规则

例如：Oracle_RVG、Exchange_RVG

SRL命名规则

例如:Oracle_SRL、Exchange_SRL

DCM磁盘命名规则

例如：Oralce_DCM1_L01_DSK, Oralce_DCM2_L20_DS

DCO磁盘命名规则

例如：Oralce_DCO1_DSK, Oralce_DCO2_DSK

软件安装与部署 — 配置管理IP地址

- 配置管理IP顺序
  - 连接终端和设备串口
  - 在终端服务器上设置串口参数，波特率为115200
  - 串口登陆，默认用户名 “admin”、密码 “Admin@storage”

```
COM1 - PuTTY

Welcome to SUSE Linux Enterprise Server 11 SP1  (x86_64) - Kernel 2.6.32.12-0.7-
default (ttyS0).

VIS_8083_0 login: admin
Password:
Login incorrect

VIS_8083_0 login: admin
Password:
1 failure since last login.
Last was Thu Oct 25 17:29:39 2012 on ttyS0.
--------------------- Welcome -----------------------
#####################################################
#                                                   #
#                                                   #
#          OceanStor VIS6000T V200R003C00           #
#                                                   #
#                                                   #
#####################################################
admin:/>
```

 Page 29  HUAWEI

软件安装与部署 — 配置管理IP地址

- 默认管理IP
  - VIS6600T机框两个控制器默认管理IP为分别为192.168.128.101和192.168.128.102
- 集群配置管理IP命令示例
  - 配置0节点控制器：

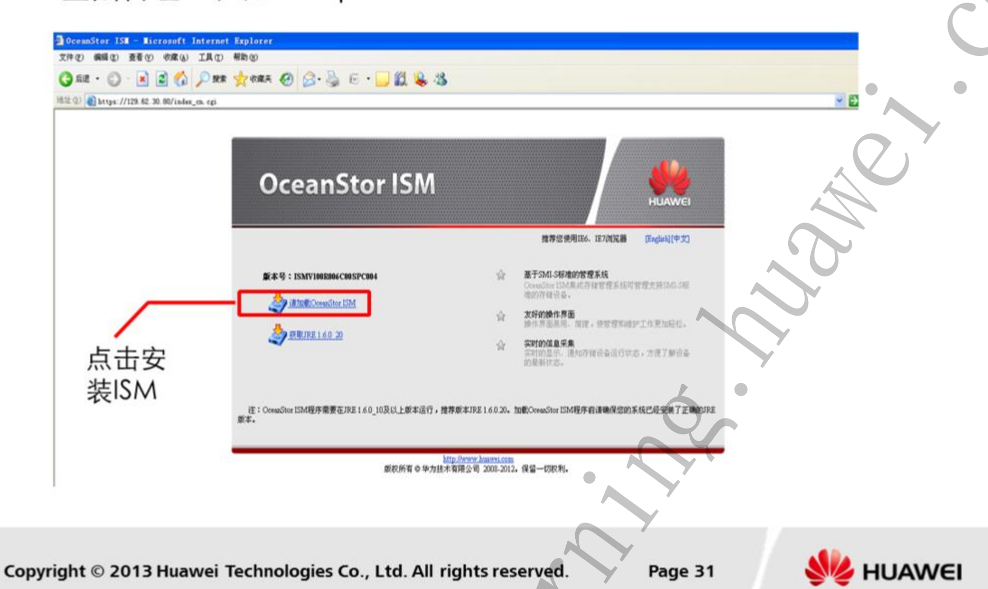
    chgctrlip －c 0 －i 2 －ip 192.168.10.10 －m 255.255.0.0 -g 192.168.0.1
  - 配置1节点控制器：

    chgctrlip －c 1 －i 2 －ip 192.168.10.11 －m 255.255.0.0 -g 192.168.0.1

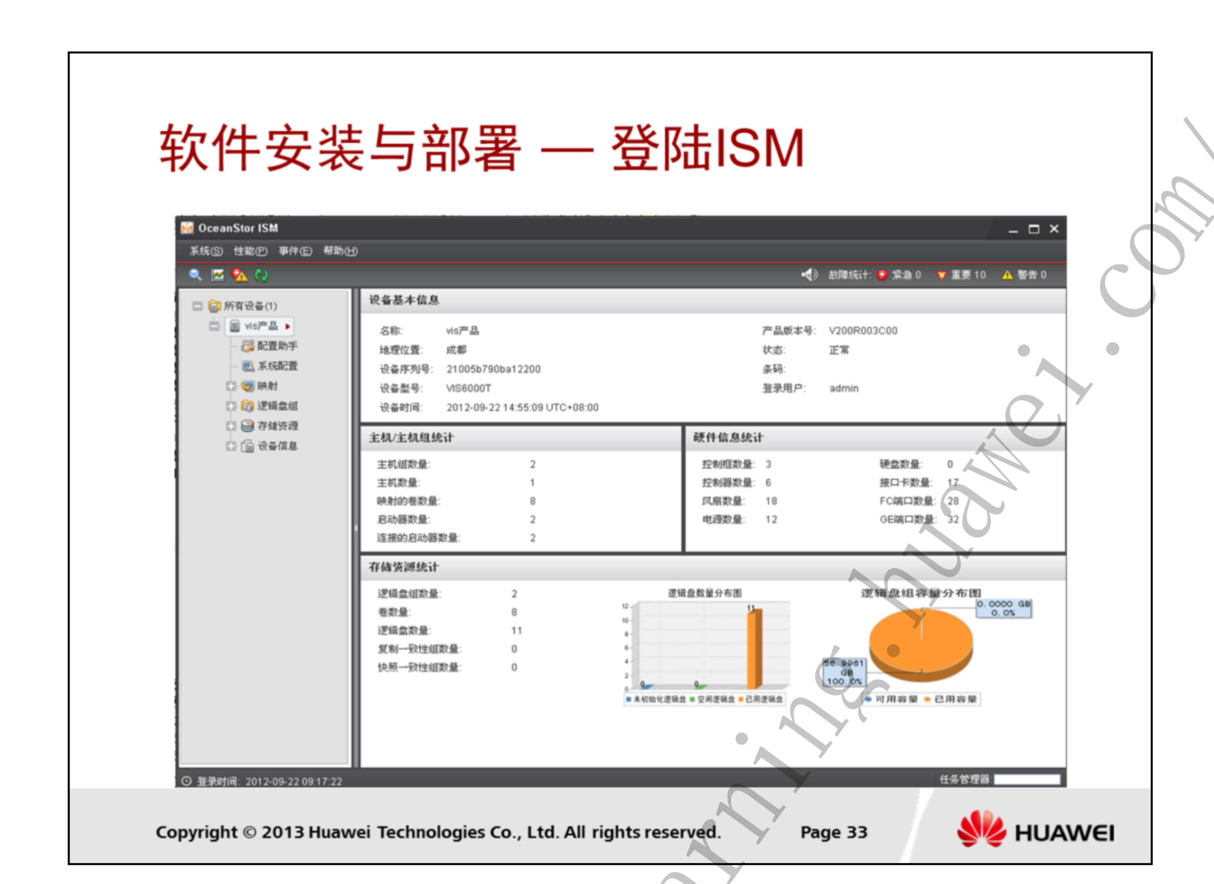


软件安装与部署 — 安装ISM

- IE登陆管理IP ， 如：https://129.62.30.80




软件安装与部署 — ISM发现设备
发现设备
请输入登录设备的用户名和密码，然后选择设备类型和发现方式。
鉴权
用户名:
admin
密码:
******
认证方式:
本设备
设备类型:
VIS
发现方式
指定IP地址 (指定设备管理网口IP地址进行发现)
IP地址:
129.62.30.80
指定IP地址段 (指定设备管理网口IP地址段进行发现)
开始IP地址:
结束IP地址:
同一子网 (在客户端所在相同子网中进行发现)
确定(O)
取消(C)
帮助(H)
设备类型
VIS
管理网口
IP地址
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 32
HUAWEI



在ISM界面可以对设备的各项运行信息进行查看，对设备运行状态进行实时监控。同时对各项功能进行配置。

软件安装与部署 — 配置集群心跳

- 步骤1 通过CLI 方式登录VIS6600T 集群的一个节点
  - 登录CLI的用户与登录GUI 的用户通用。初始时，请使用VIS6600T 提供的用户名为“admin”、**密码为**“Admin@storage”的超级管理员用户登录。
- 步骤2 使用chgheartmode 命令设置集群心跳模式

<u>注：集群超过2个节点，必须配置为外部心跳。</u>

Page 34

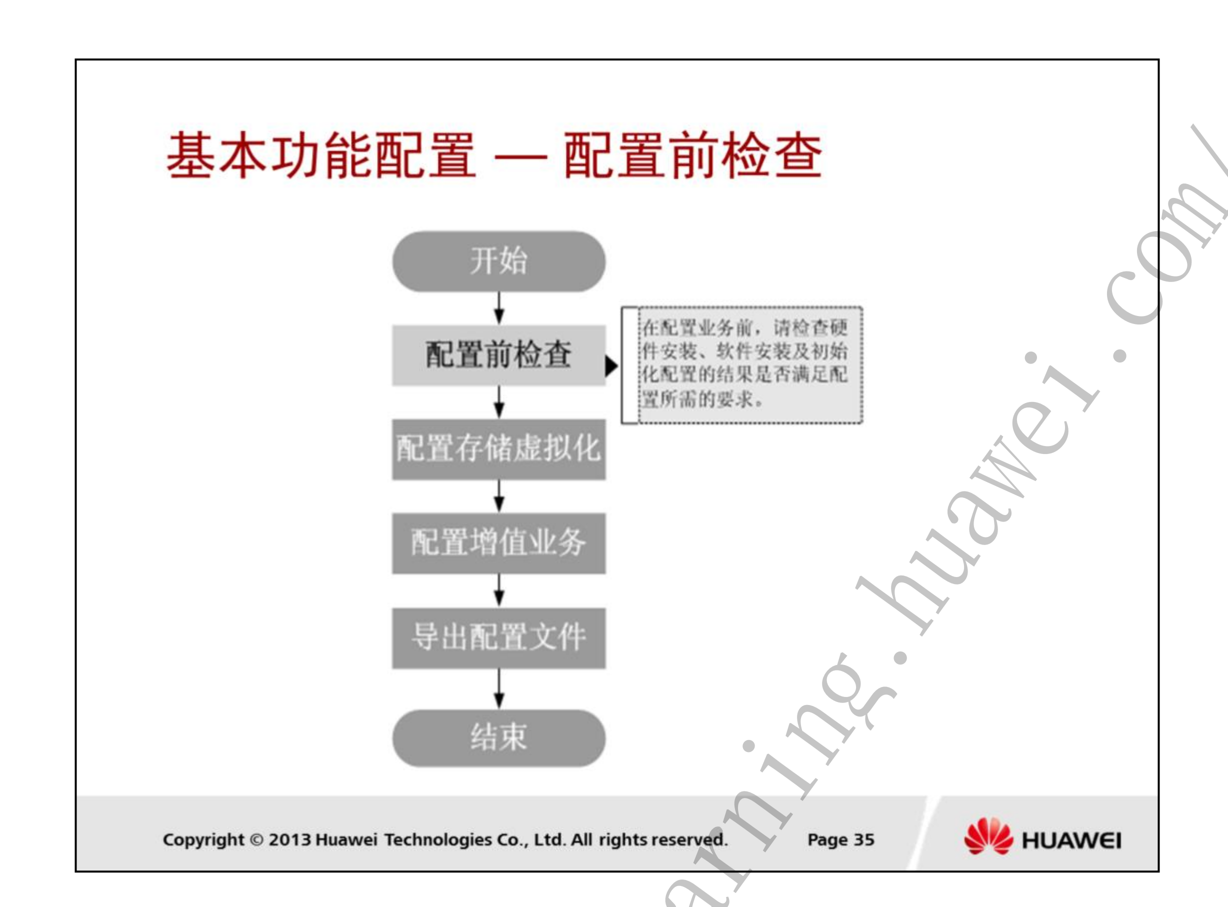


chgheartmode 命令的格式和参数说明如下：

chgheartmode -m *mode*

mode为集群的心跳模式，0——内部心跳；1——外部心跳。

命令执行时会提示确认修改，确认输入y，回车后集群自动重启。


基本功能配置 — 配置前检查
开始
配置前检查
在配置业务前，请检查硬件安装、软件安装及初始化配置的结果是否满足配置所需的要求。
配置存储虚拟化
配置增值业务
导出配置文件
结束
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 35
HUAWEI

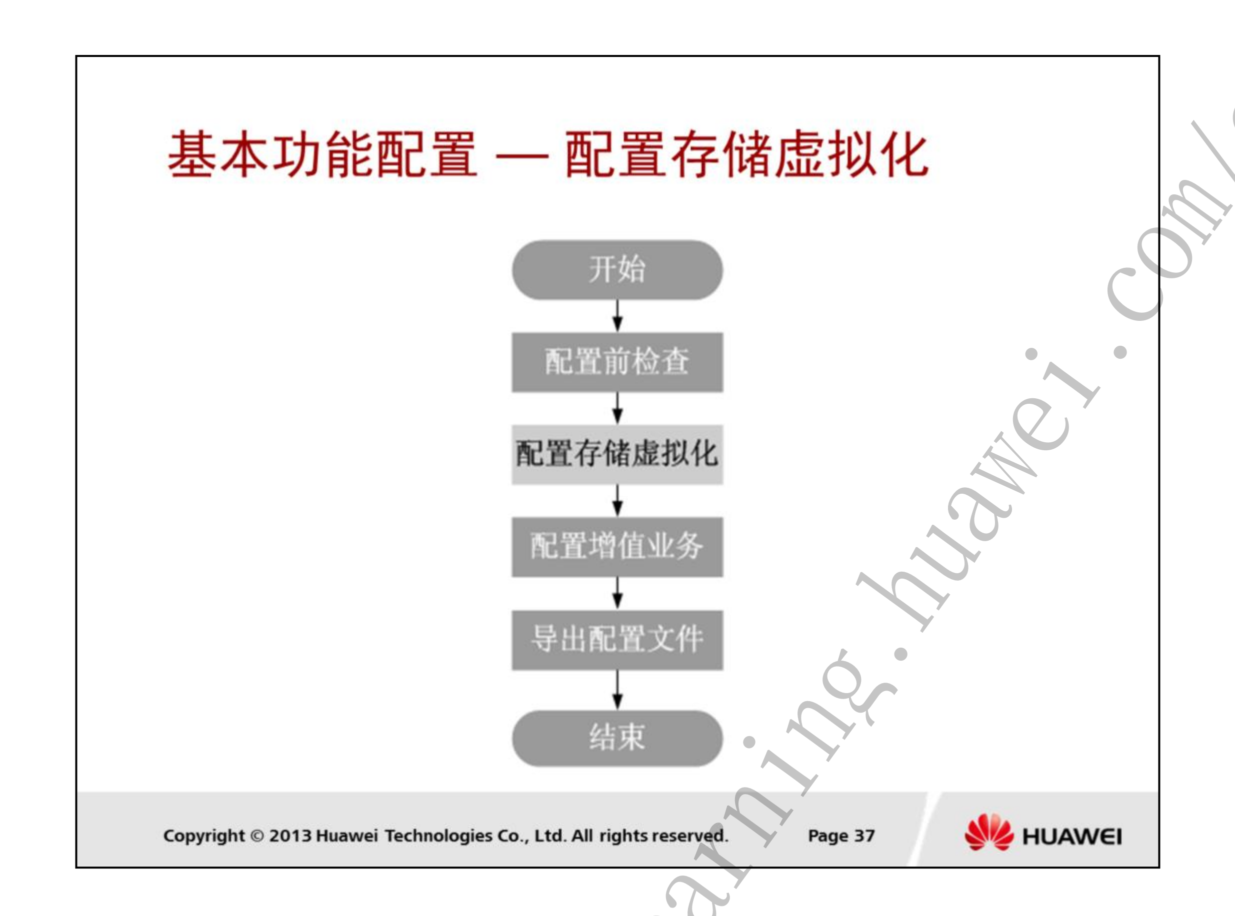
基本功能配置 — 配置前检查

- 检查硬件安装
- 检查软件安装
- 检查初始化配置结果
- License 文件

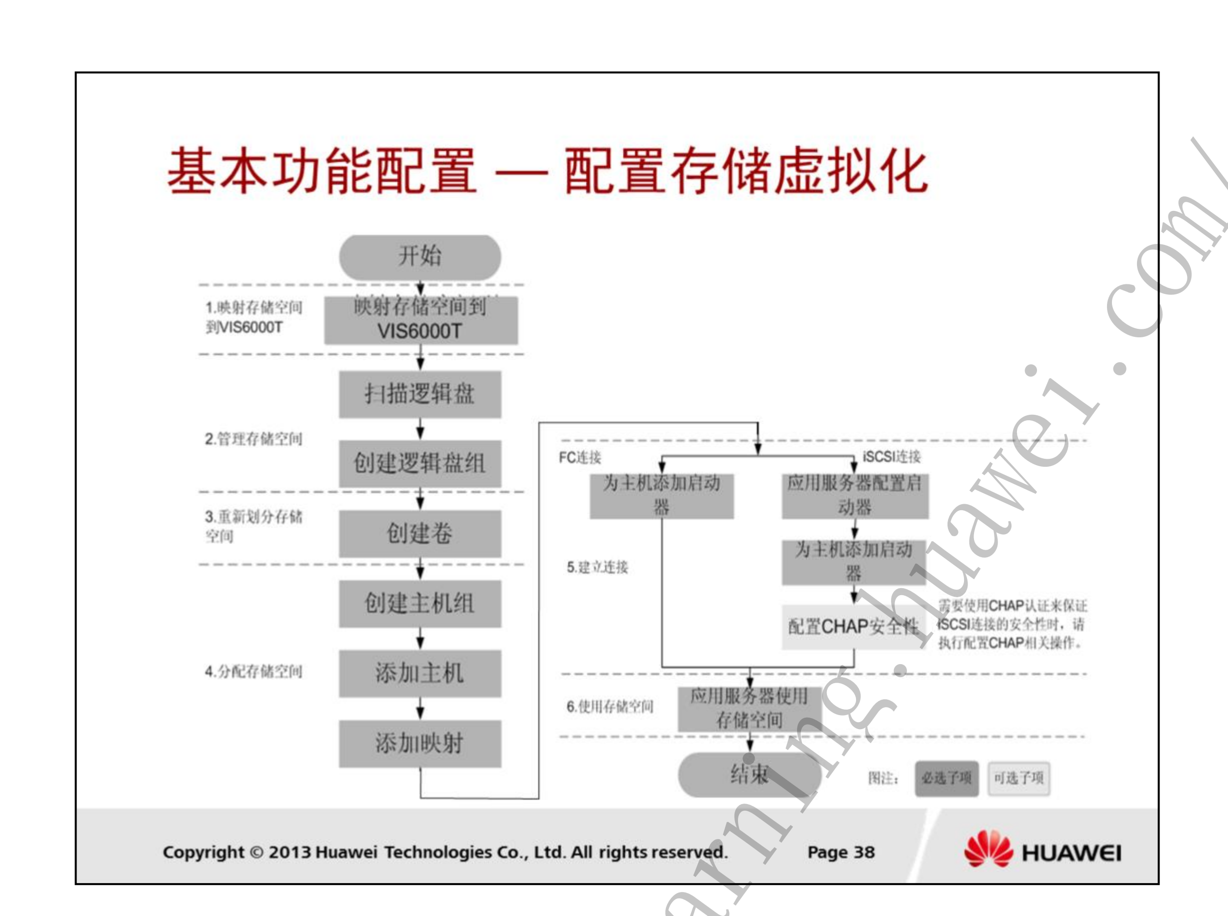
 Page 36



- 检查软件安装
    - VIS6600T 侧软件，PC机安装ISM并成功发现设备。
    - 服务器软件，相应系统的iSCSI启动器以及多路径软件。
- 检查初始化配置结果
    - 维护终端与VIS6600T之间的连接
    - 阵列与VIS6600T之间的连接
    - VIS6600T与应用服务器之间的连接

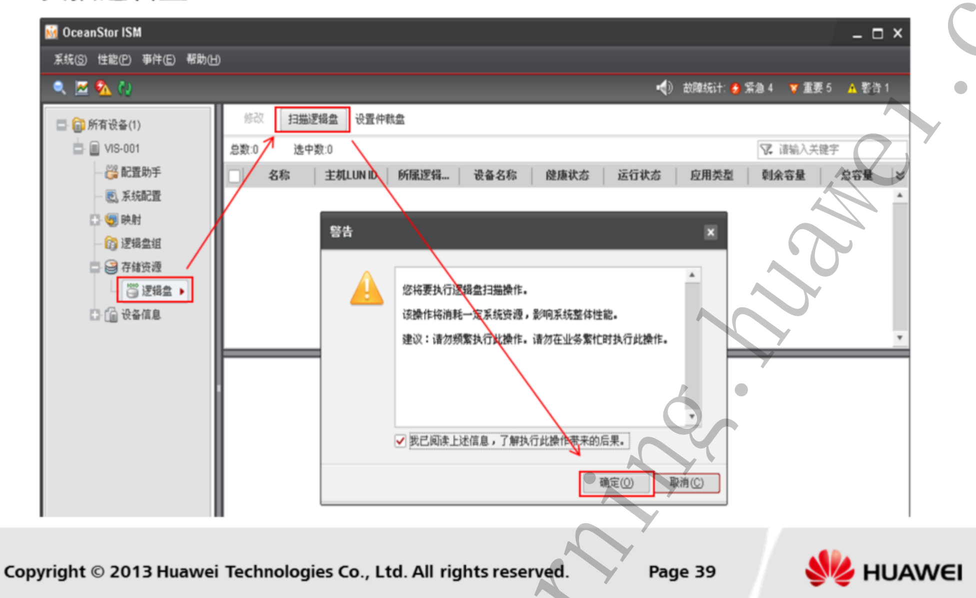


存储虚拟化是VIS最重要的功能，是配置所有高级功能的前提。

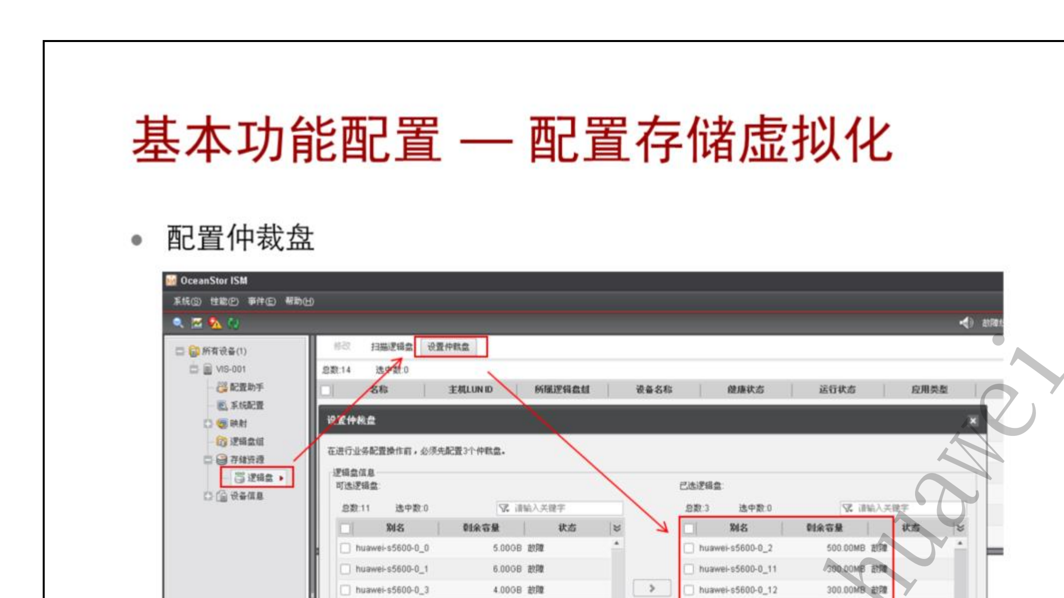

基本功能配置 — 配置存储虚拟化
开始
1.映射存储空间到VIS6000T
映射存储空间到VIS6000T
扫描逻辑盘
2.管理存储空间
创建逻辑盘组
3.重新划分存储空间
创建卷
创建主机组
4.分配存储空间
添加主机
添加映射
FC连接
iSCSI连接
为主机添加启动器
应用服务器配置启动器
5.建立连接
为主机添加启动器
配置CHAP安全性
需要使用CHAP认证来保证iSCSI连接的安全性时，请执行配置CHAP相关操作。
6.使用存储空间
应用服务器使用存储空间
结束
图注：
必选子项
可选子项
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 38
HUAWEI

基本功能配置 — 配置存储虚拟化

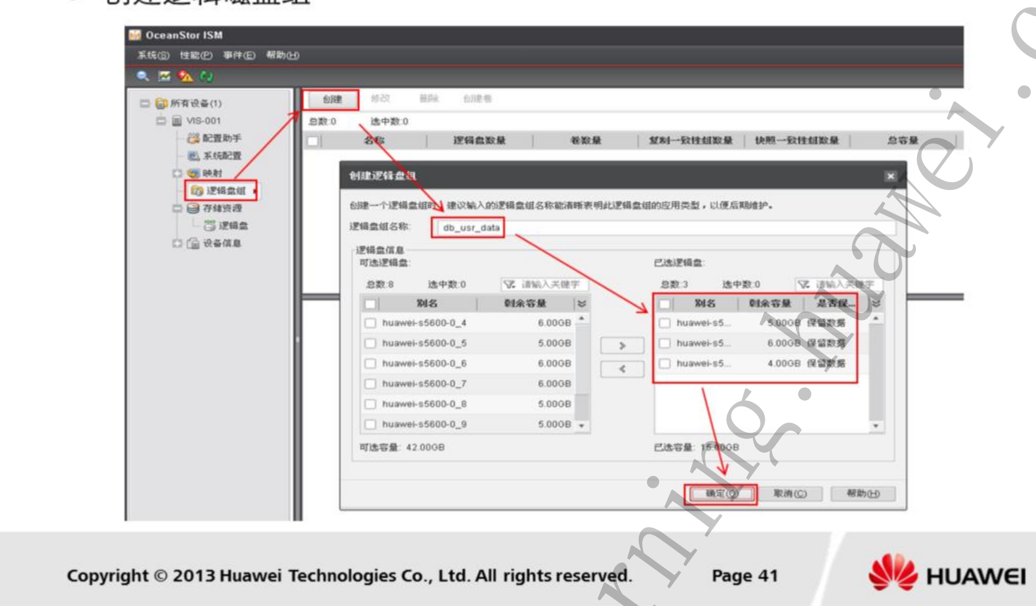
- 扫描逻辑盘



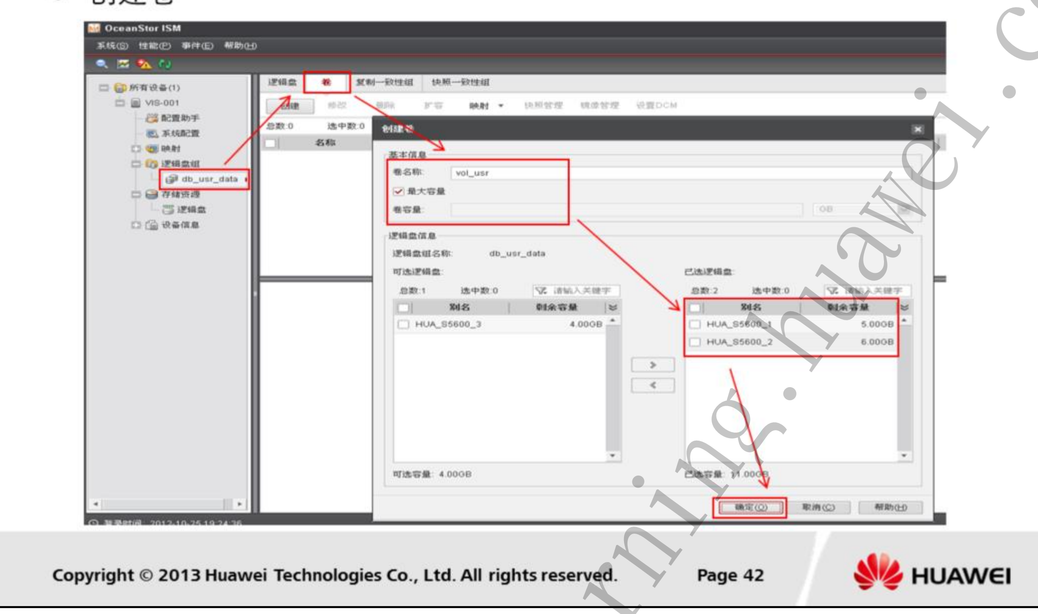
在存储资源中的逻辑盘选项中选择扫描逻辑盘



在存储资源中的逻辑盘选项中选择设置仲裁盘，仲裁盘要设置3个。



在映射中选择逻辑盘组选项，再点击创建，选择所需的逻辑盘加入到新创建的逻辑盘组中。



在逻辑盘组中选择已经创建好的逻辑盘组，在逻辑盘组选项中点击卷，再选择创建，在创建新卷时可以设定卷名，卷容量和卷所使用的逻辑盘。



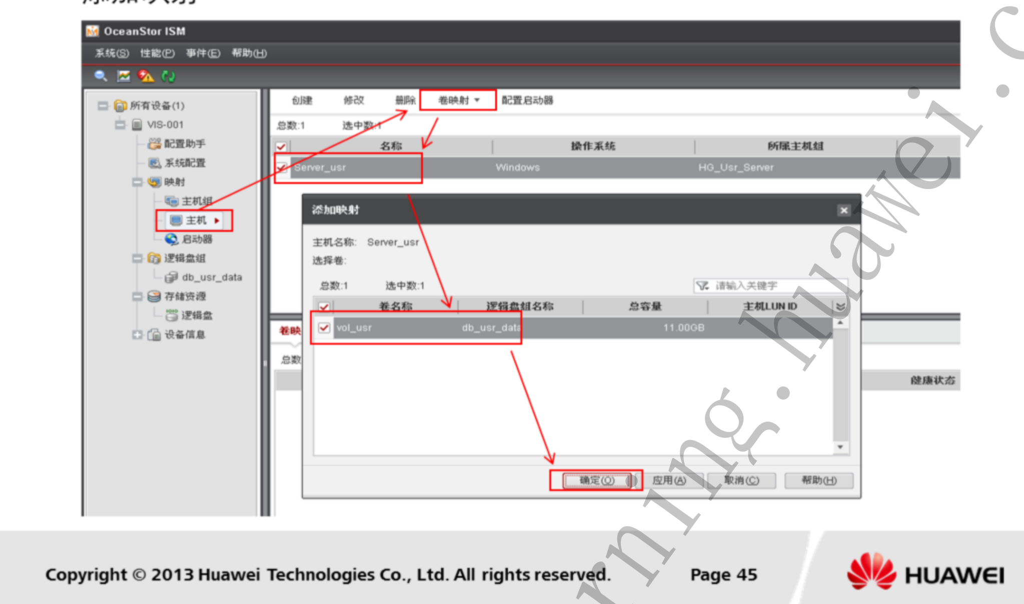
在映射选项中选择主机组，并点击创建来创建新的主机组，创建主机组时只需要设置主机组名。



在映射中选择主机，点击创建引出创建主机指引，根据指引可逐步设置主机名，主机所属主机组，主机操作系统，主机的FC启动器或iSCSI启动器。

基本功能配置 — 配置存储虚拟化

- 添加映射

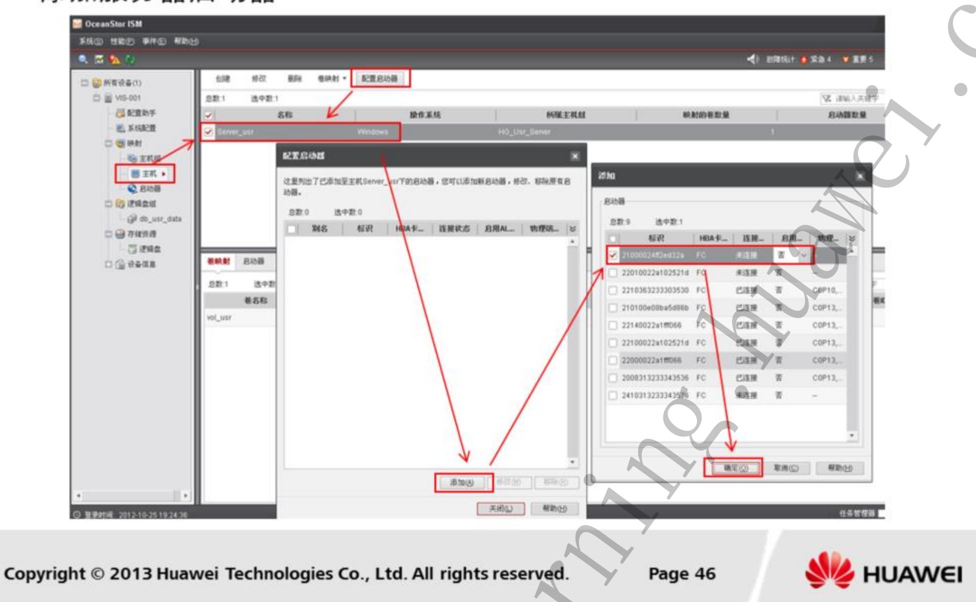


 Page 45

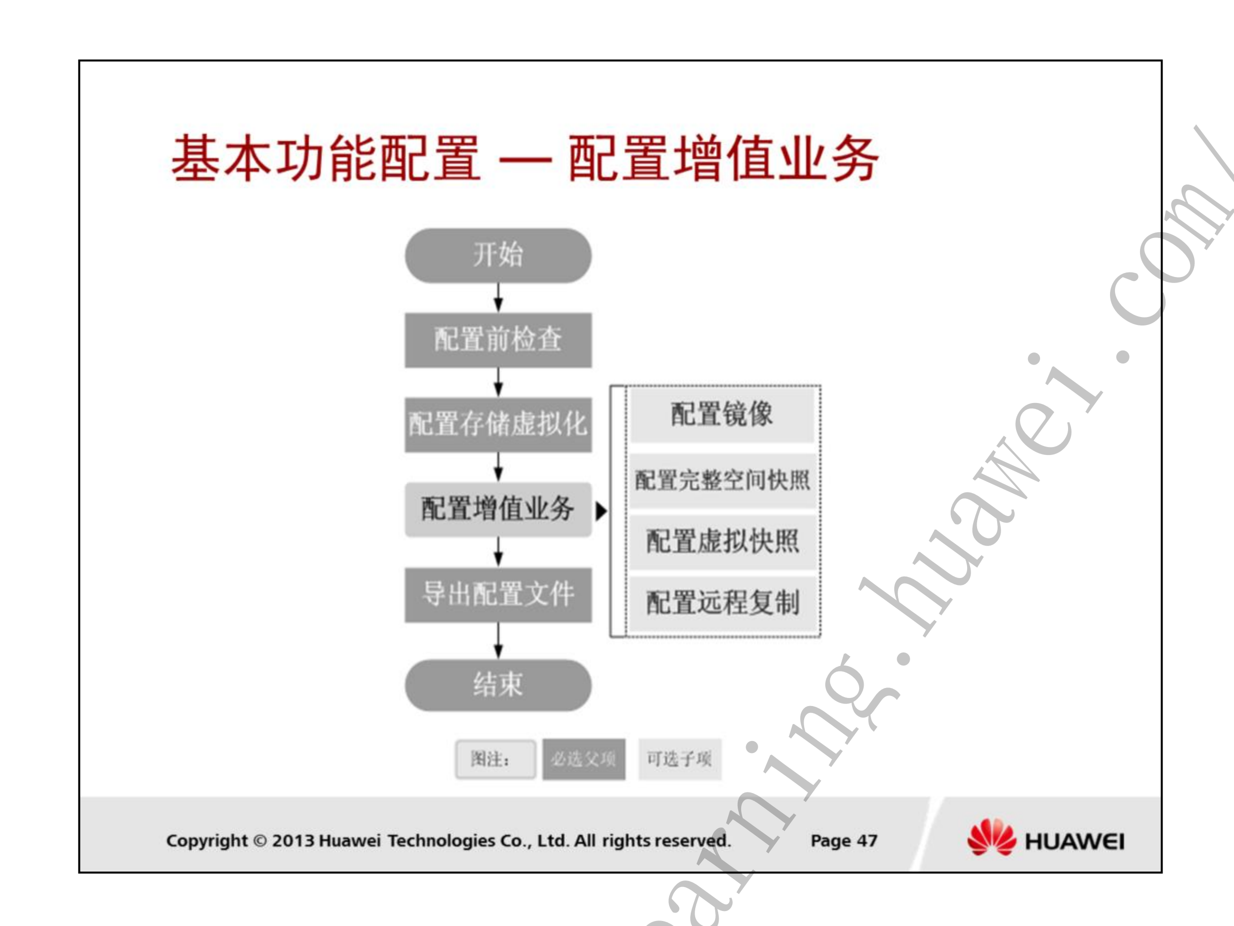
在映射下选择主机，选择需要添加映射的服务器，点击卷映射后选择相应的卷映射给主机。

基本功能配置 — 配置存储虚拟化

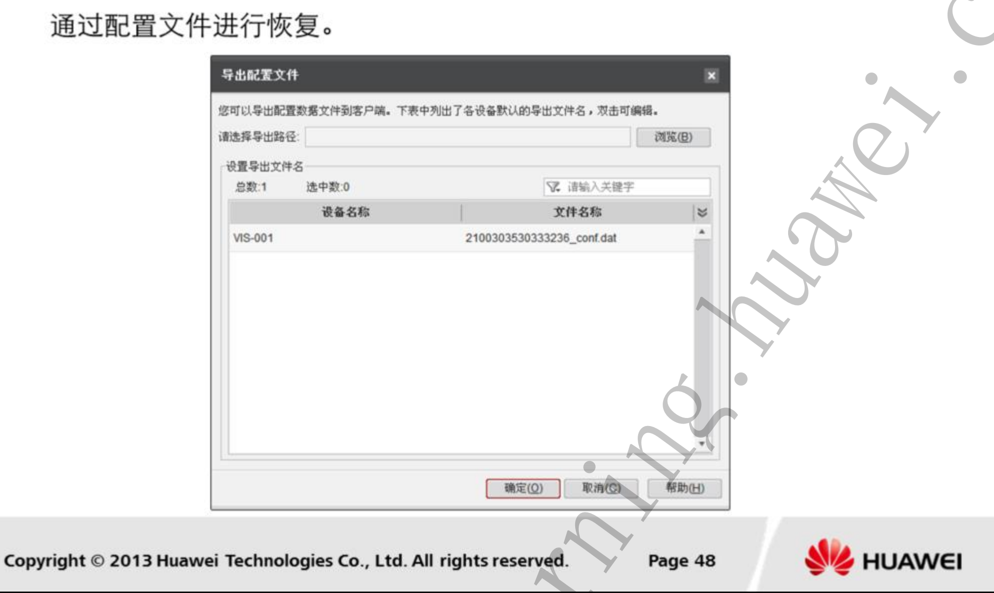
- 添加服务器启动器



在映射下选择主机，选择需要手动配置启动器的主机，点击配置启动器，选择添加，在列表中选择服务器对应的启动器添加给服务器。



VIS增值业务包含：镜像，完整空间快照，虚拟快照和远程复制。在存储虚拟化配置完成后，可以根据业务需要配置相应的功能。



步骤1：进入“导出配置文件”对话框

在导航树上选择“所有设备 > VIS6600T > 系统配置”。在右侧的信息展示区的“导入导出”区域单击“导出配置文件”。

步骤2：设置配置文件的保存路径及名称

单击“浏览”。在“保存位置”下拉列表中选择配置文件的保存路径。或者在“文件夹名”文本框中输入配置文件的保存路径。单击“保存”。

步骤3：（可选）修改导出的配置文件名称

双击设备名称后对应的文件名称，然后输入新的文件名称（扩展名为.dat）。

步骤4：保存配置文件

单击“确定”。系统弹出“提示”对话框。请仔细阅读“提示”对话框中的内容，确认后单击“确定”。系统弹出“执行结果”对话框，提示配置文件导出的结果。如果系统提示操作成功，在选择的路径下可以查看文件名为“*.dat”的配置数据。如果系统提示保存失败，重新输入配置文件的文件名再进行导出操作。

步骤5：单击“关闭”，完成配置文件的导出。

基本功能配置 — 导出配置文件

- 对设备的所有设置都会被保存在系统配置文件中，配置文件包括逻辑盘组、卷、iSCSI主机端口IP地址、FC主机端口速率等的配置信息，一旦损坏或丢失，会造成业务中断甚至数据丢失的风险。
- 为了保证系统的安全性和可靠性，在对设备进行了初始配置或者重要的配置更改后，都建议对配置文件进行导出保存，以便在系统由于客观因素或人为因素发生重大故障造成配置信息失效或丢失时使用备份的配置文件对系统进行恢复。





目录



VIS6600T产品完整空间快照功能原理

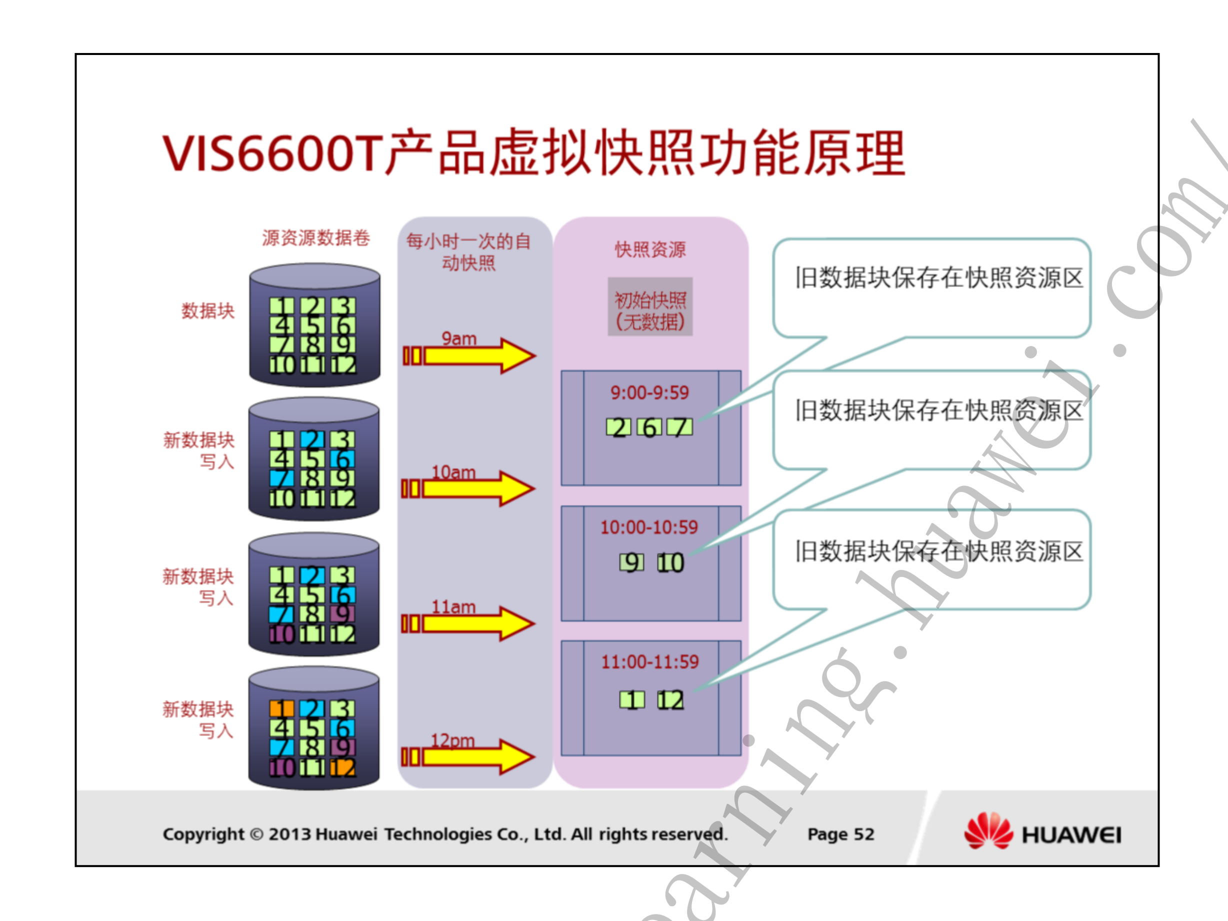
- 完整空间快照技术的实现原理：在快照时间点到来时，系统会为源数据卷分配一个大小完全相同的物理空间作为快照卷，并启动后台数据同步，在同步数据完成后，该时间点快照创建成功。



 Page 51 HUAWEI



1. 创建一个跟源卷大小一致的卷作为快照卷，并开始后台数据同步。
2. 在数据同步过程中如果源卷有新数据写入，写入的数据位置为还未没有同步拷贝的内容，则将原数据写入到快照卷中，新数据写入源卷，保持源卷数据为最新状态；如写入的数据位置为同步拷贝完成的部分，则只将新数据写入源卷；快照卷数据内容不变。
3. 在数据全部同步完成后，快照卷与9：00的源卷数据完全相同，此时快照结束。



在VIS6600T快照创建前，数据写入动作与没开启快照功能的VIS6600T一样：对数据进行修改，会直接写入原有磁盘区块对原有数据进行覆盖，原有数据不会被保留。

在创建虚拟快照之前，需要进行虚拟快照的相应配置，首先创建快照资源池，。快照资源池由若干资源卷组成，当资源池中的容量不足以支撑时，可以动态的向资源池中添加新的空间。推荐每个源卷用一个快照资源池，快照资源池的空间必须为本磁盘组内的空间，推荐为同一RAID组内空间；同时需要创建DCO（Data Change Object）用来记录数据改变位图、快照资源池数据和快照之间的映射关系等。

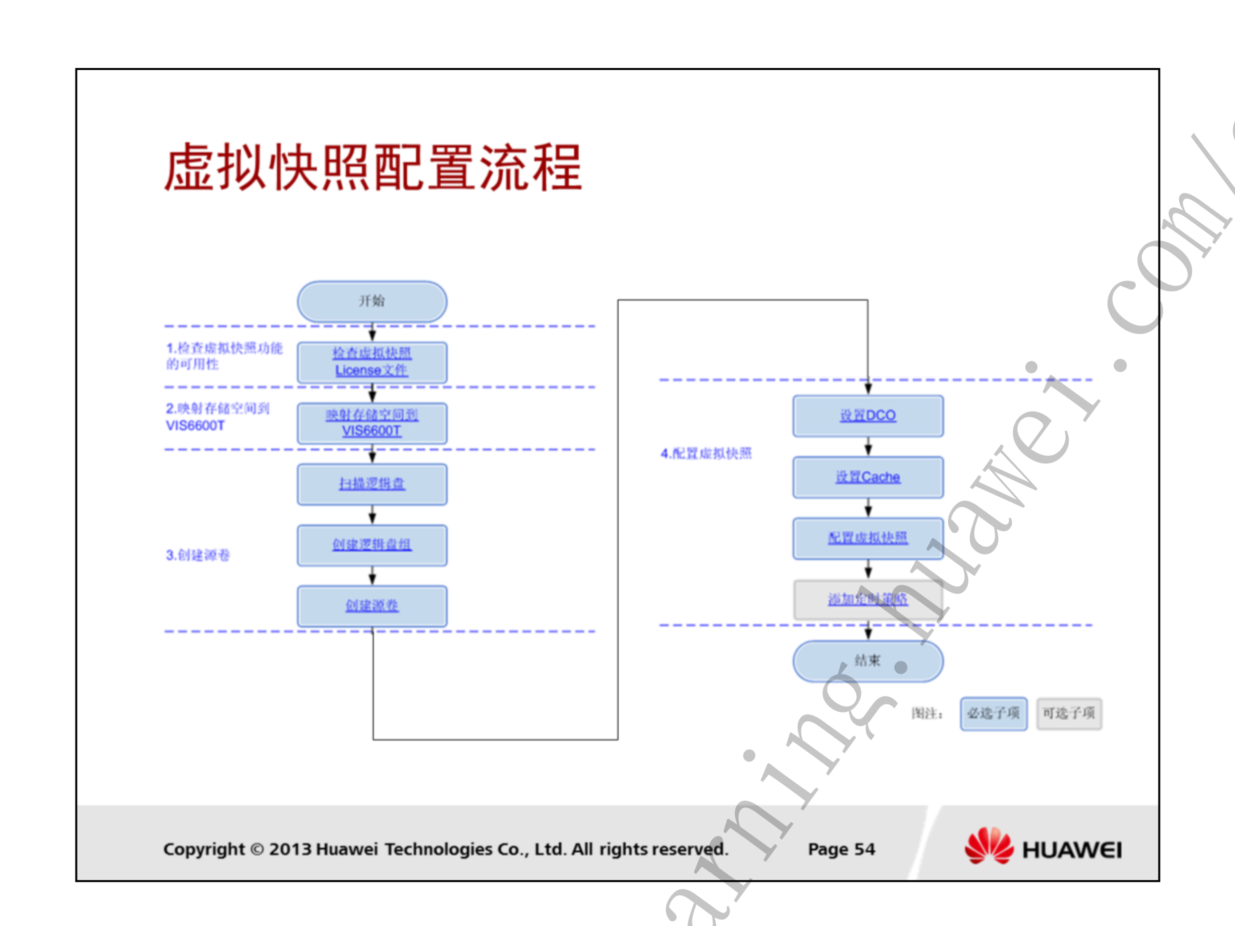
当虚拟快照被创建时，系统同时创建一个映射表和位图。位图记录数据改变块的位置，映射表记录了源卷中原数据和其物理地址的映射关系，而快照资源池将用于保存源卷各个磁盘区块在快照时间点后第一次被更新时源卷相应位置的原数据，在一个快照周期内无论再对原磁盘区块的数据进行怎样的更新，当前的数据均不再被保存到资源池，而是被直接覆盖。

虚拟快照相关概念

概念	描述
源卷	用作创建快照的卷被称之为源卷。
一致性组	多个卷在业务上相互联系，需要作为一个实体进行统一管理，这个实体就是一致性组。
DCO（Data Change Object）盘/DCO镜像盘	DCO盘和DCO镜像盘用于创建DCO。DCO中记录的是映射关系改变的相关位图，而改变的数据块是存储在对应的Cache中。通过DCO中的映射关系，可以对快照执行读写、刷新、回滚等操作。
上水位	Cache自动增长的水位，如果I/O超过了上水位，则Cache将自动增长。
增值粒度	Cache自动增长的大小。
写前拷贝	某一时间点的虚拟快照创建并激活后，当有数据写入源卷时，首先将需要更新数据的位置的原数据移动到Cache中，同时映射表修改映射关系，记录原数据的新位置，然后再将新数据写入到源卷中。此过程即为写前拷贝（copy-on-write）技术。
映射表	用于记录数据和其存储的物理地址映射关系的表。映射表中的左项为源卷中数据的地址，作为查找键值；右项为Cache中数据的地址。
资源池	用于虚拟快照过程中，保存源卷改变数据的存储空间，由一个或多个普通的卷组成，这类卷被称为“资源池”。

 Page 53





虚拟快照配置流程
开始
1.检查虚拟快照功能的可用性
检查虚拟快照License文件
2.映射存储空间到VIS6600T
映射存储空间到VIS6600T
扫描逻辑盘
3.创建源卷
创建逻辑盘组
创建源卷
4.配置虚拟快照
设置DCO
设置Cache
配置虚拟快照
添加定时策略
结束
图注：
必选子项
可选子项
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 54
HUAWEI

VIS6600T产品快照典型组网与连接

同时支持FC、iSCSI，以FC组网为例：

VIS6600T产品快照功能应用规划

- DCO规划
  - 推荐用于创建DCO 的DCO 盘和DCO 镜像盘分别来源于不同的阵列
  - DCO 盘和DCO 镜像盘基本为写操作，其性能需求比对应的源卷或是目标卷高1 ～2 倍，推荐来源于写性能好的RAID10
- Cache规划
  - 公式计算

    Cache 的大小 = 源卷在虚拟快照刷新周期内的数据变化量 * 虚拟快照的个数 * 120%
  - 推荐值

    在无法确定数据变化量的情况下，推荐Cache的大小与源卷相同

 Page 56 

- DCO盘和DCO镜像盘是指用来存放数据更改对象的逻辑盘。DCO镜像盘与DCO盘互为镜像关系。
- 当在同一个逻辑盘组中创建不同的增值业务（虚拟快照、完成空间快照和镜像）时，推荐使用同一组DCO盘和DCO镜像盘。
- 用作DCO盘和DCO镜像盘的逻辑盘需要设置为“不保留数据”的逻辑盘。
- 设置DCO时，请确认至少有2个空闲逻辑盘。

虚拟快照配置参数

- 创建虚拟快照需设置DCO和Cache
- DCO盘和DCO镜像盘是指用来存放数据更改对象的逻辑盘。DCO镜像盘与DCO盘互为镜像关系。
- Cache用于存储源数据相对于某个时间点的变化，同一个源卷的所有虚拟快照共用同一个Cache。

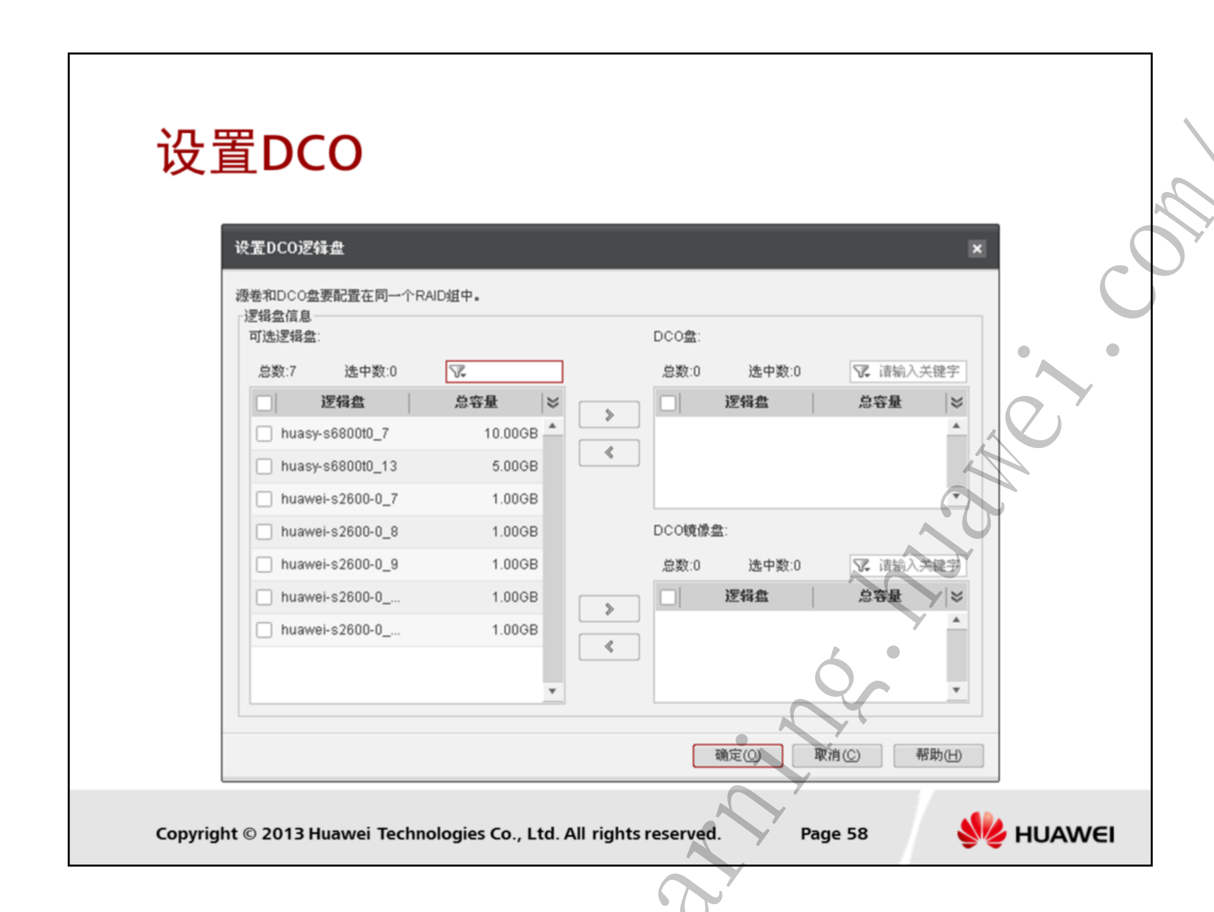


Page 57
HUAWEI

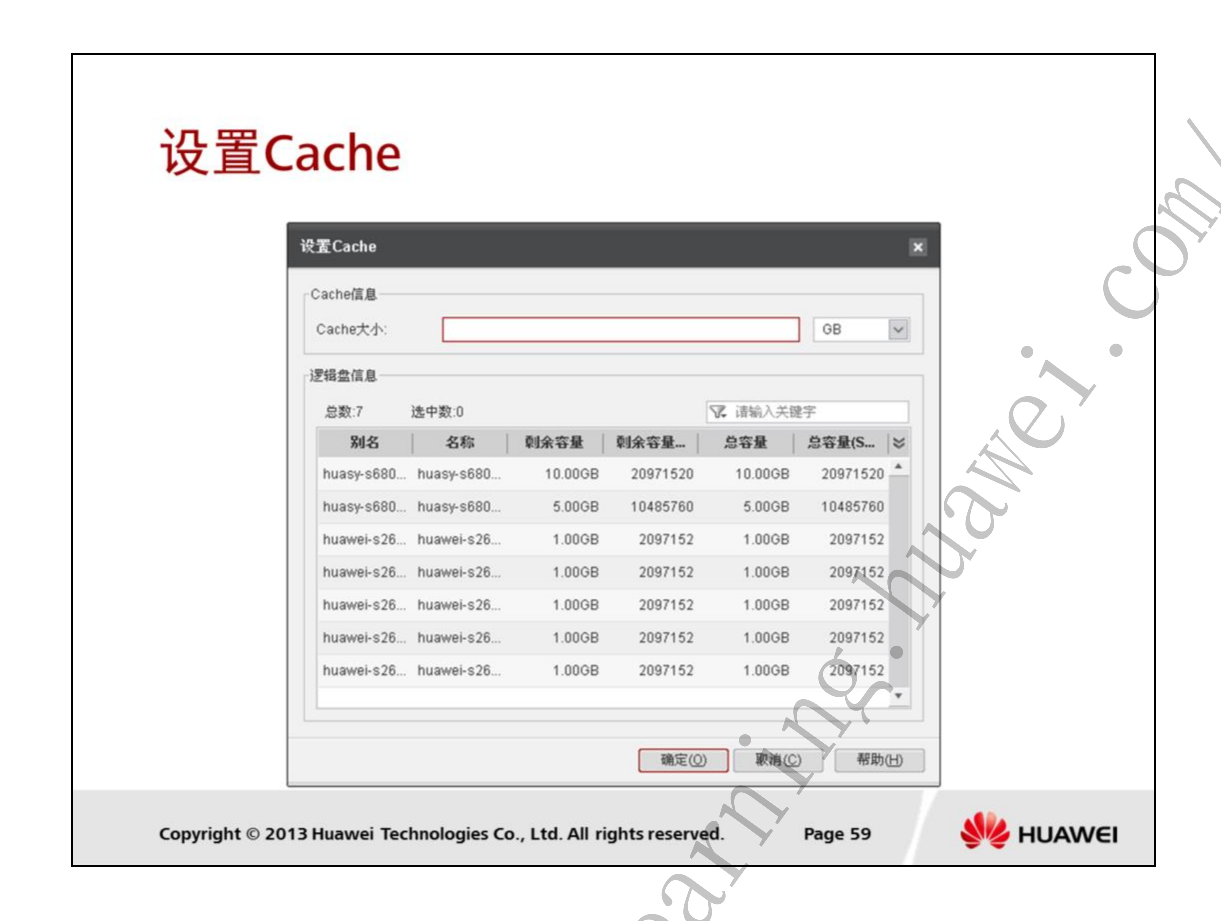
当在同一个逻辑盘组中创建不同的增值业务（虚拟快照、完整空间快照和镜像）时，推荐使用同一组DCO盘和DCO镜像盘,用于设置DCO的2个逻辑盘在加入逻辑盘组时，必须设置为“不保留数据”的逻辑盘。

虚拟快照不需要源卷的存储空间的完整副本，而是使用Cache。每当源卷的原始内容将要被覆盖时，会在执行写入之前将源卷中的原始数据复制到Cache中。在创建虚拟快照时可以配置该Cache的大小，所需存储空间比源卷少得多。Cache的存储空间默认可以自动增长，当Cache的存储空间不足时，可以使用逻辑盘组中任何可用空闲空间自动增加Cache的大小。虚拟快照提高了写入性能，并且需要的存储空间也较少。

Cache容量推荐值：源卷在虚拟快照刷新周期内的数据变化量*虚拟快照的个数*120%。例如，为源卷datavol创建了7个虚拟快照，虚拟快照的定时刷新策略为每天刷新一次，则Cache容量为：datavol一天内的数据变化量*7*120%。



从可选逻辑盘列表中选择相同大小的两个盘分别放置在DCO盘和DCO镜像盘选项下。

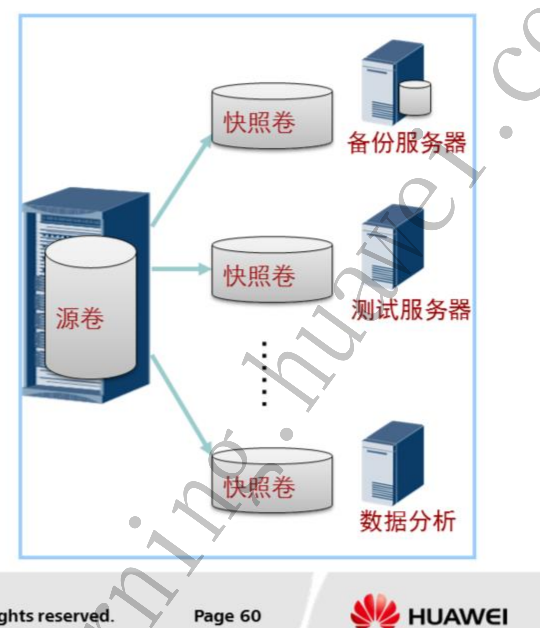


Cache用于存储源数据相对于某个时间点的变化，同一个源卷的所有虚拟快照共用同一个Cache。

Cache容量推荐值：源卷在虚拟快照刷新周期内的数据变化量*虚拟快照的个数*120%。例如，为源卷datavol创建了7个虚拟快照，虚拟快照的定时刷新策略为每天刷新一次，则Cache容量为：datavol一天内的数据变化量*7*120%。

VIS6600T产品快照功能应用

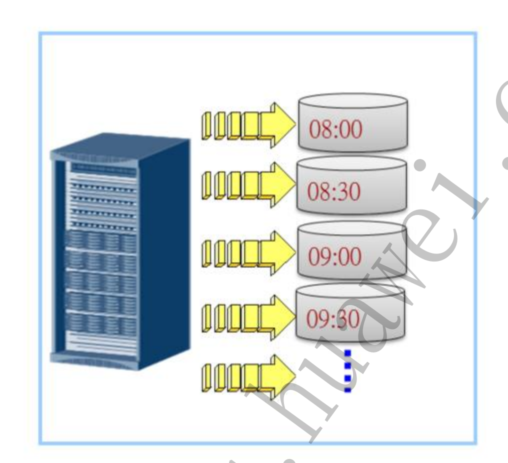
- 场景一：快照数据用于并行处理
  - 在几乎不影响源卷业务的情况下生成源卷数据的快照
  - 同一源卷可以生成多份快照





VIS6600T产品快照功能应用

- 场景二：持续数据保护
  - 实现对数据卷的持续保护，能恢复到指定的时间点的数据。
  - 虚拟快照仅保存变化数据，最大限度减少存储空间占用。

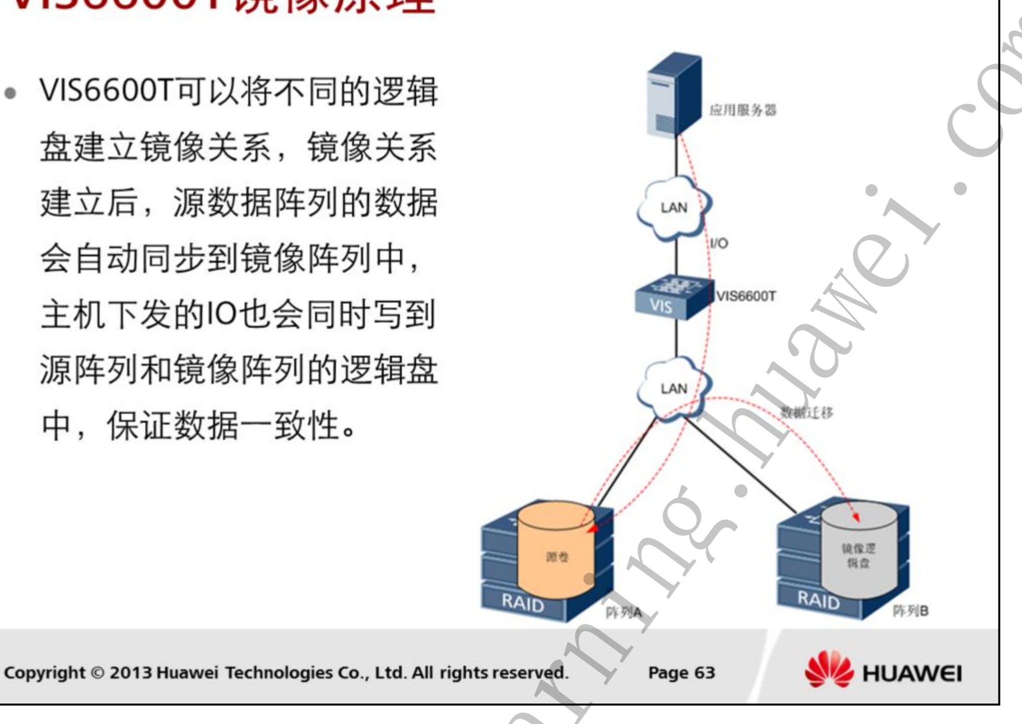






目录

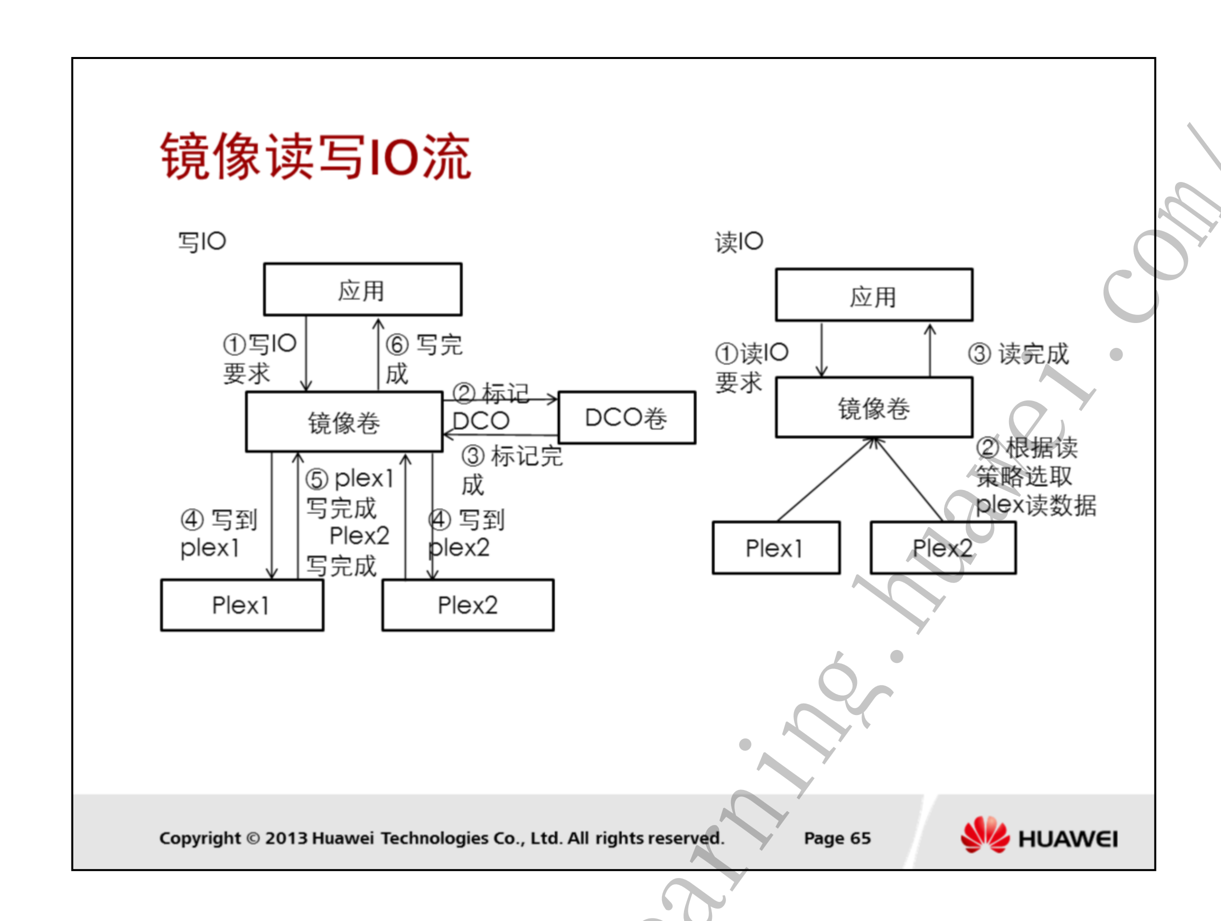



VIS6600T镜像原理
VIS6600T可以将不同的逻辑盘建立镜像关系，镜像关系建立后，源数据阵列的数据会自动同步到镜像阵列中，主机下发的IO也会同时写到源阵列和镜像阵列的逻辑盘中，保证数据一致性。
应用服务器
LAN
I/O
VIS
VIS6600T
LAN
数据迁移
源卷
RAID
阵列A
镜像逻辑盘
RAID
阵列B
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 63
HUAWEI

镜像相关概念

概念	描述
镜像	镜像特性通过数据副本提供数据冗余。每个副本（或镜像）都存储在与此卷不同的逻辑盘上。对一个卷创建镜像后，可确保在组成镜像的任意一方发生故障时，其数据不会丢失，提高了数据的完整性和可靠性。
源卷	用于创建镜像的卷。
镜像逻辑盘	用于与源卷创建镜像关系，与源卷保存同样的数据，容量必须大于等于源卷。推荐源卷与镜像逻辑盘分别来源于不同的存储设备，且源卷与镜像逻辑盘的RAID（Redundant Array of Independent Disks）级别相同。
DCO（Data Change Object）盘/DCO镜像盘	DCO盘和DCO镜像盘用于创建DCO。DCO用于管理快速重同步的映射信息，与某个卷关联，则可以对该卷使用快速重同步功能。例如：互为镜像关系的源卷和镜像逻辑盘，当源卷故障时，则从源卷故障时刻开始，与卷关联的DCO会通过位图的方式记录镜像逻辑盘的修改数据。当源卷修复后，源卷可以根据DCO记录的差异信息同步镜像逻辑盘中修改的数据，这样可以达到快速重同步的效果。
镜像的读策略	轮循模式：非连续空间的I/O，以轮流循环方式将读访问请求下发到不同的存储空间，实现存储空间的并发响应，提升系统性能。 优先模式：即主备模式，所有的读访问I/O请求都默认下发给优先的存储空间响应，当优先空间故障时才下发给备选存储空间，故障修复后切换回优先空间。





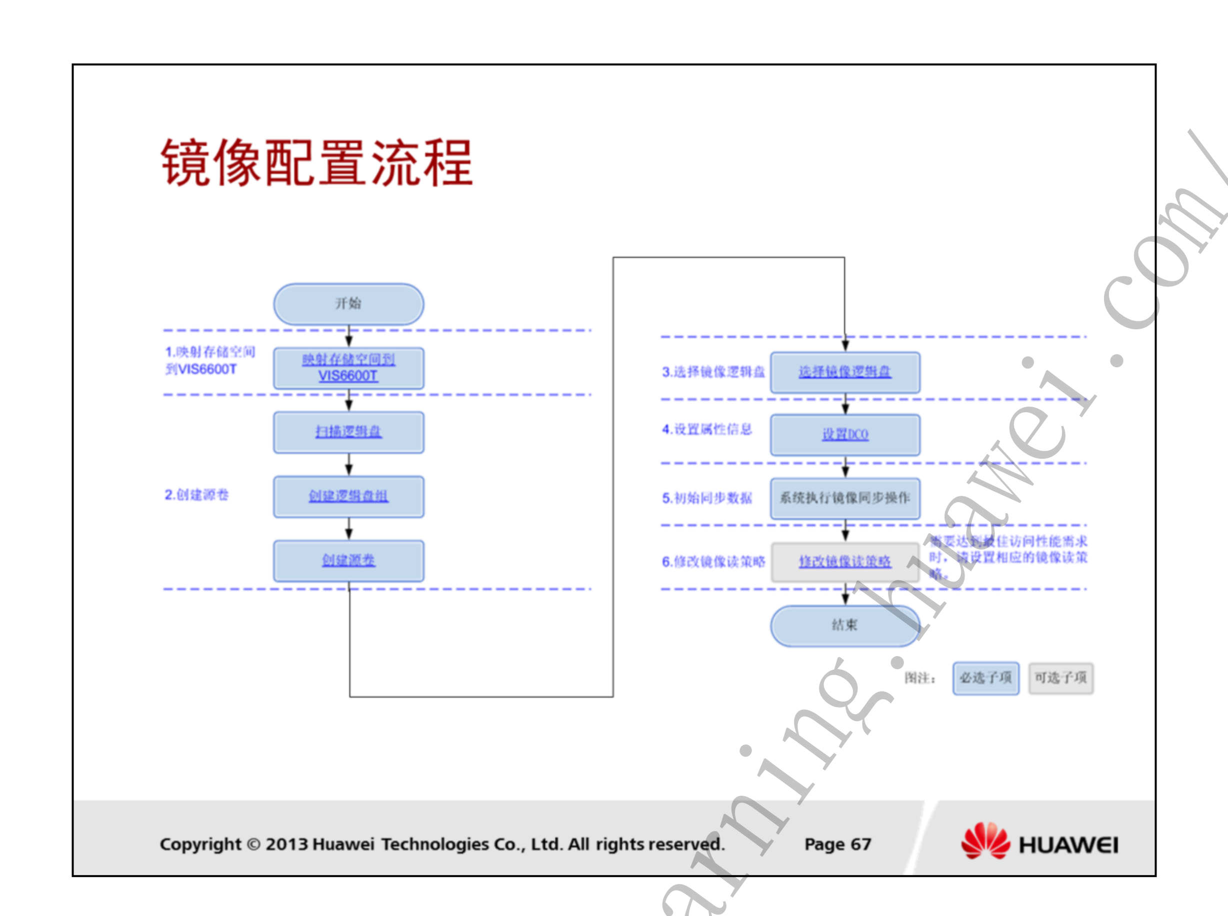
- 读策略：
  - 优先:首先从已命名为prefer的 plex 中读取。
  - 轮循:对检测到的读 I/O，以 “循环” (round-robin) 模式依次读取每个 plex。

镜像同步

- 镜像同步过程中，下发I/O的处理

需要同步的数据	处理方式
未同步的数据	2个plex同时下发I/O，确保数据一致性
已同步的数据	2个plex同时下发I/O，确保数据一致性
正在同步的数据	等同步完成后，2个plex同时下发I/O，确保数据一致性




镜像配置流程
开始
1.映射存储空间到VIS6600T
映射存储空间到VIS6600T
扫描逻辑盘
2.创建源卷
创建逻辑盘组
创建源卷
3.选择镜像逻辑盘
选择镜像逻辑盘
4.设置属性信息
设置DCO
5.初始同步数据
系统执行镜像同步操作
6.修改镜像读策略
修改镜像读策略
需要达到最佳访问性能需求时，请设置相应的镜像读策略。
结束
图注：
必选子项
可选子项
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 67
HUAWEI

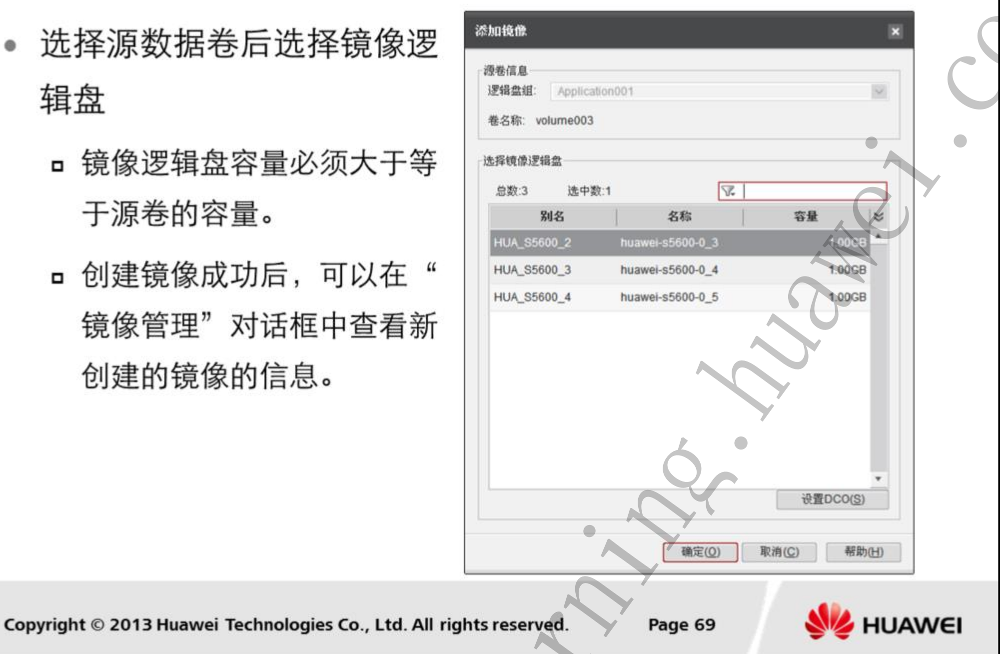
VIS6600T镜像功能应用规划

- 数据卷
  - 源卷：一般来自客户现有系统中应用的LUN；
  - 镜像卷：空间来自镜像用阵列，和源盘不在同一个阵列；大于或等于源盘的容量，精确粒度为sector；性能高于或等于源盘。
- DCO
  - 源DCO：推荐和数据卷的源盘来自同一个阵列；
  - 镜像DCO：和数据卷的镜像盘来自同一个阵列，不能和源DCO盘在同一个阵列上。
- 数据盘读策略
  - 本地镜像：轮询（缺省）；
  - 同城镜像：优先

 Page 68



- 优先:首先从已命名为prefer的 plex 中读取.
- 轮循:对检测到的读 I/O，以“循环” (round-robin) 模式依次读取每个 plex.
- 一组镜像中，源DCO和镜像DCO配置大小推荐值均为1G；如有多组镜像，则相应增加。


创建镜像
选择源数据卷后选择镜像逻辑盘
镜像逻辑盘容量必须大于等于源卷的容量。
创建镜像成功后，可以在“镜像管理”对话框中查看新创建的镜像的信息。
添加镜像
源卷信息
逻辑盘组:
Application001
卷名称: volume003
选择镜像逻辑盘
总数:3
选中数:1
别名
名称
容量
HUA_S5600_2
huawei-s5600-0_3
1.00GB
HUA_S5600_3
huawei-s5600-0_4
1.00GB
HUA_S5600_4
huawei-s5600-0_5
1.00GB
设置DCO(S)
确定(O)
取消(C)
帮助(H)
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 69
HUAWEI

设置DCO

- DCO用于管理快速重同步的映射信息，与某个卷关联，则可以对该卷使用快速重同步功能。



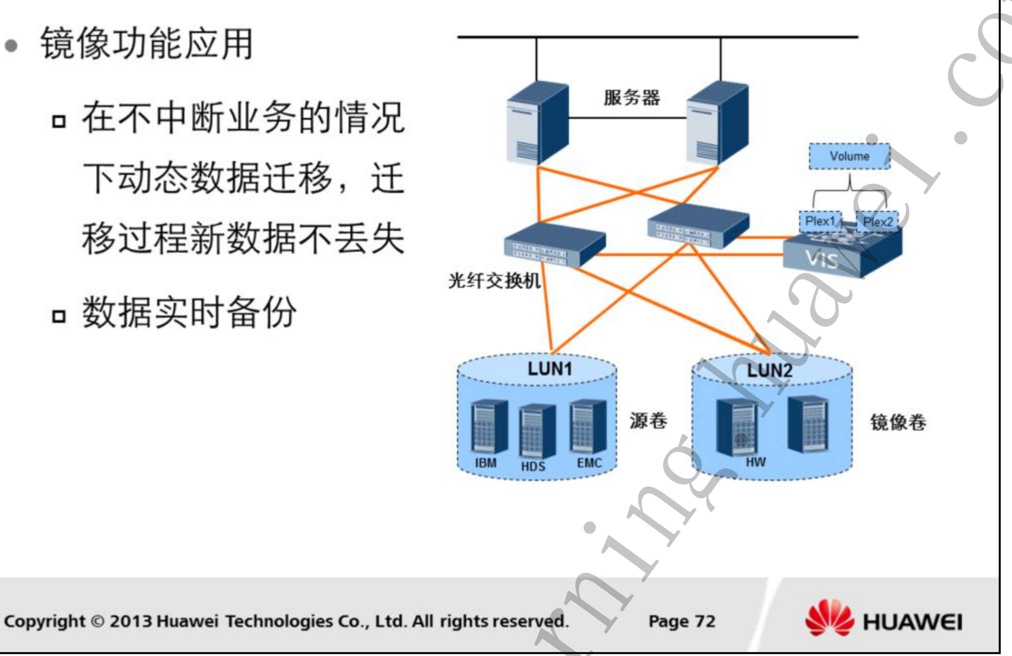
DCO的作用是记录镜像关系的数据差异。逻辑盘A与逻辑盘B为镜像关系，当逻辑盘A故障时，则从逻辑盘A故障时刻开始，作为DCO的逻辑盘就会通过位图的方式记录逻辑盘B的修改数据。当逻辑盘A修复后，逻辑盘A可以根据DCO记录的差异信息同步逻辑盘B中修改的数据，这样可以达到快速重同步的效果。

当在同一个逻辑盘组中创建不同的增值业务（虚拟快照、完整空间快照和镜像）时，推荐使用同一组DCO盘和DCO镜像盘。

用于设置DCO的2个逻辑盘在加入逻辑盘组时，必须设置为“不保留数据”的逻辑盘。



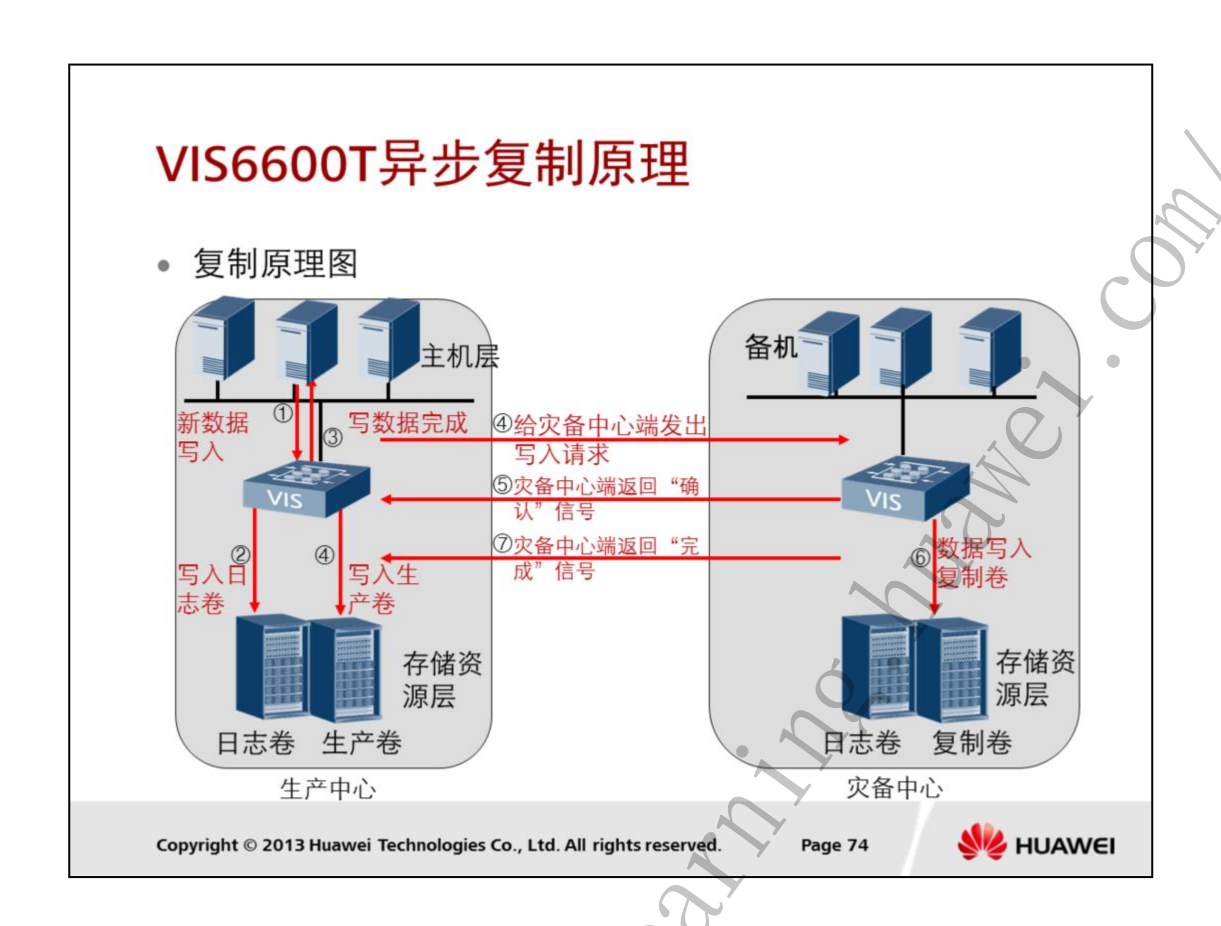
- 轮循模式：非连续空间的I/O，以轮流循环方式将读访问请求下发到不同的存储空间，实现存储空间的并发响应，提升系统性能。在镜像创建成功后，镜像的读策略缺省为轮循模式。
- 优先模式：即主备模式，所有的读访问I/O请求都默认下发给优先的存储空间响应，当优先空间故障时才下发给备选存储空间，优先控制的故障修复后切换回优先空间。


VIS6600T镜像功能应用
镜像功能应用
在不中断业务的情况下动态数据迁移，迁移过程新数据不丢失
数据实时备份
服务器
Volume
Plex1
Plex2
VIS
光纤交换机
LUN1
LUN2
IBM
HDS
EMC
源卷
HW
镜像卷
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 72
HUAWEI



目录





异步远程复制，生产中心写入数据后直接反馈主机写完成，生产中心到灾备中心的写入操作后续完成，因而灾备中心和生产中心的数据存在一定的时延。

远程复制相关概念

概念	描述
复制一致性组	VIS6600T的远程复制解决方案是通过复制一致性组实现的。逻辑盘组中需要配置异地容灾的卷组，这些卷组由一个或多个相关卷组成，可以同时复制到灾备中心，实现数据备份，例如数据库。
虚拟IP地址	每个复制一致性组的数据复制地址。生产中心和灾备中心应该各自有一个唯一的虚拟IP地址，且能相互通信。
数据卷	在生产中心用于存储应用数据，是需要通过复制一致组特性创建异地容灾的卷。
复制卷	在灾备中心用于存储应用数据副本的卷。
日志卷	复制一致性组数据写入缓冲区，该缓冲区具有循环写入的功能。用于跟踪记录复制一致组中每个数据卷的数据写入顺序的卷，每次数据依次写入日志卷和数据卷。为提高数据的写入性能，建议将数据卷和日志卷创建在不同的物理硬盘上。每个复制一致性组均需配置一个日志卷。
DCM	全称为Data Change Map，当日志卷溢出时，跟踪生产中心的数据卷写入顺序的卷，由DCM卷和DCM镜像卷组成。当生产中心和灾备中心的链路恢复正常时，将生产中心未复制到灾备中心的数据重新同步到灾备中心。
生产中心	业务系统日常运行的环境。在复制一致性组中，生产中心的VIS6600T将把数据复制到灾备中心，形成数据副本。
灾备中心	业务系统日常备用的环境。正常情况下，灾备中心的VIS6600T实时将来自生产中心的数据写入到本地的复制卷中。当生产中心发生计划内或计划外事件时，灾备中心将立即启用该数据副本，保证业务连续性。

 Page 75



容灾切换

当生产中心需要计划内维护时，可以在生产中心启用容灾切换功能，将业务系统切换到灾备中心，确保业务连续性。

灾备接管

当生产中心出现无法恢复的故障时，可以在灾备中心启用灾备接管功能，将业务系统接管到灾备中心，确保业务连续性。

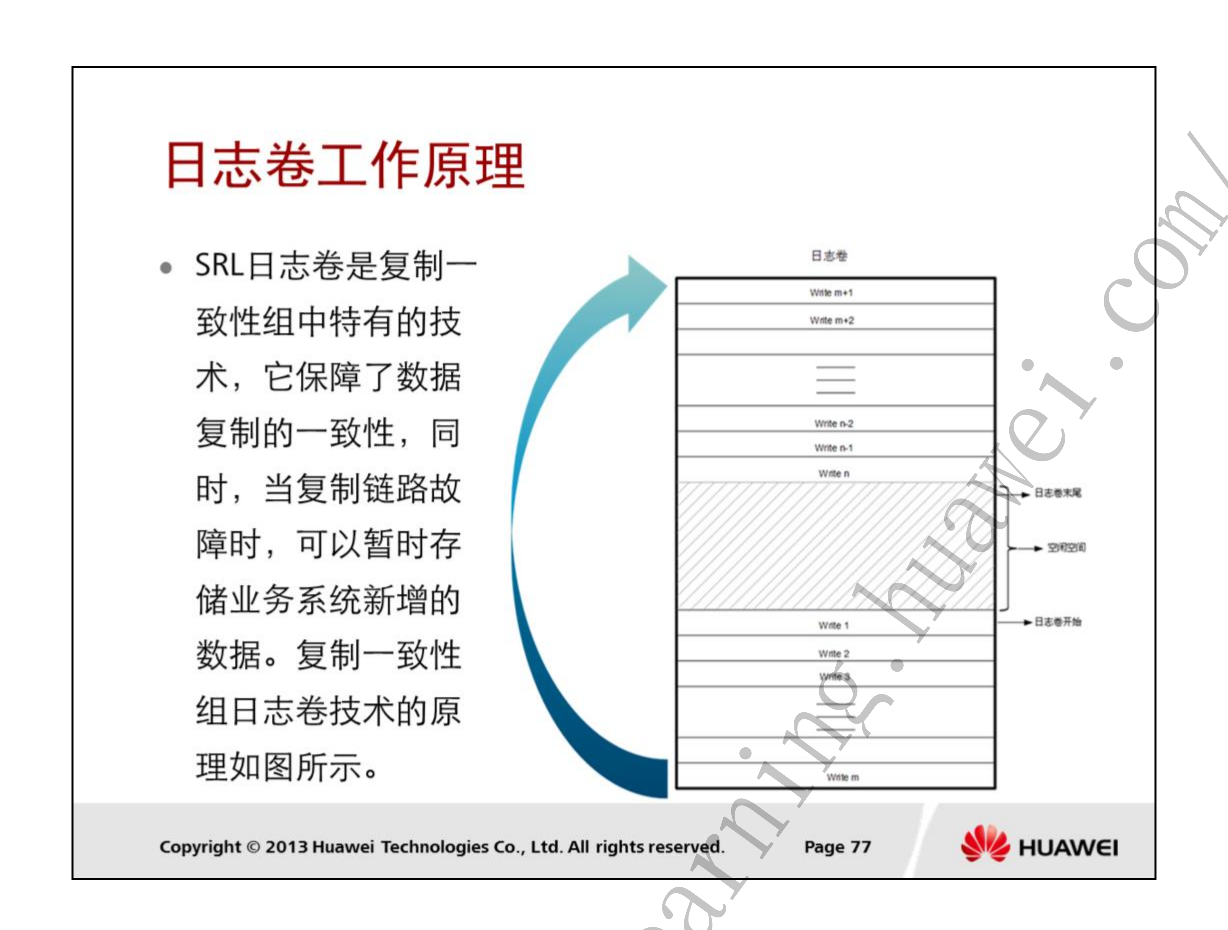
灾备回切

当生产中心恢复后，可以在灾备中心启用灾备回切功能，将灾备接管期间产生的新数据复制到生产中心。

VIS6600T复制原理-复制组术语介绍

- SRL（Storage Replicator Log）：生产灾备中心都需要创建，用于保证生产的数据按照主机下发I/O的顺序复制到灾备，保证数据一致性；
- DCM (Data Change Map)：每个复制数据卷都需要关联，通常为互为镜像的两个LUN；DCM主要用于SRL溢出后记录生产和灾备数据变化，在恢复复制时避免采用全同步；
- RVG(Replicated Volume Group):复制卷组是给定 VxVM 磁盘组内为复制配置的卷组。RVG 始终是 VxVM 磁盘组的子集。可以将磁盘组中的一个或多个相关卷配置为 RVG。相关卷是指一组必须在辅助节点上将应用程序写入按序复制到其中的卷。





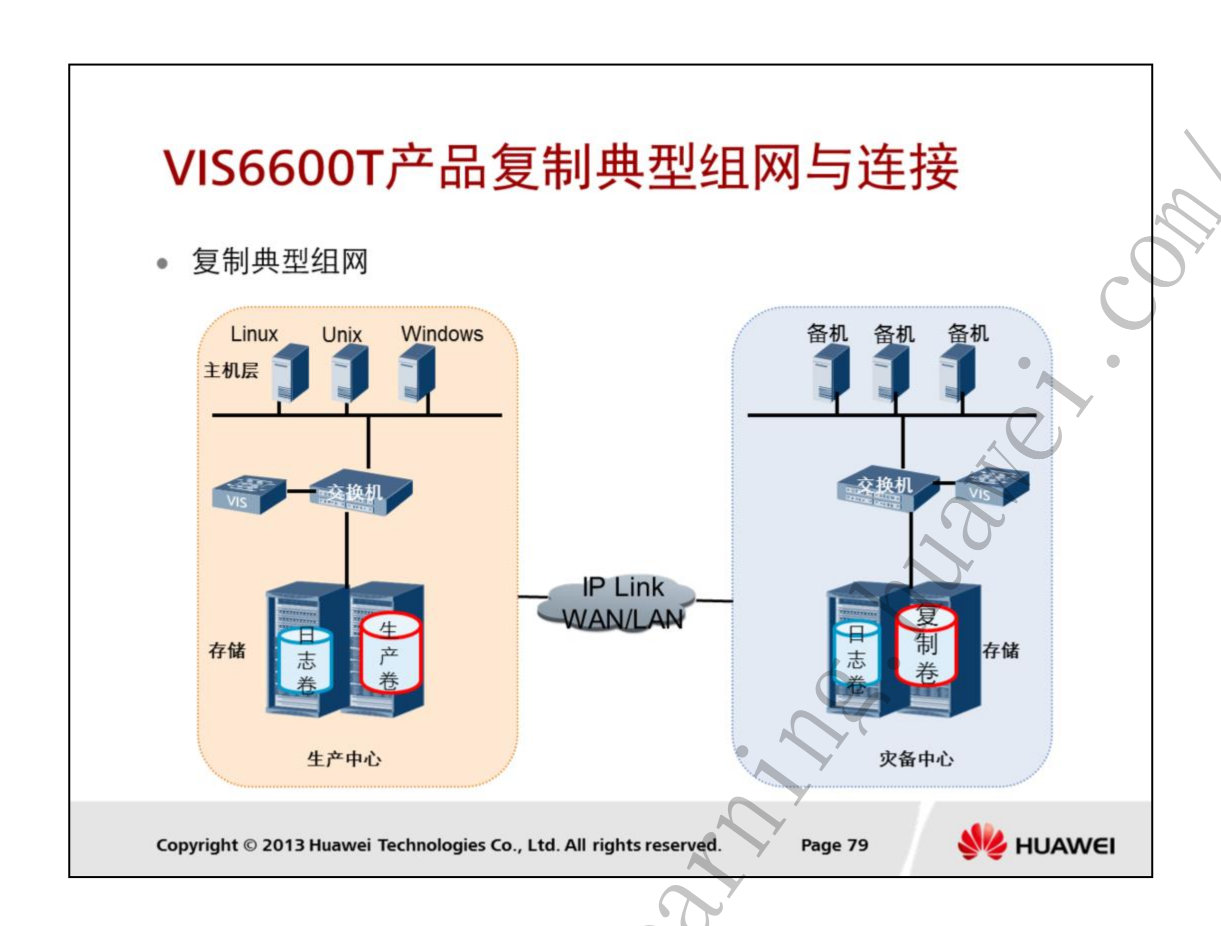
当第1个写I/O达到队列，日志卷从开始处开始记录。接着第2个I/O，第3个I/O，直至第N个I/O写到了日志卷的最末端。当N+1个I/O到达时，将重新从日志卷的开始处重新记录。写入日志卷的I/O将按照日志卷记录的顺序，将其写入到生产中心和灾备中心的数据卷中存储。当生产中心收到灾备中心发来的写I/O收到的确认信息后，VIS6600T在日志卷中将该写I/O的状态标记为完成，实现生产中心和灾备中心数据的一致性的同时，将该I/O从日志卷中清除，将该I/O暂用的日志卷空间留给后达到的I/O，从而实现了日志卷空间的循环使用。

VIS6600T产品复制功能关键技术

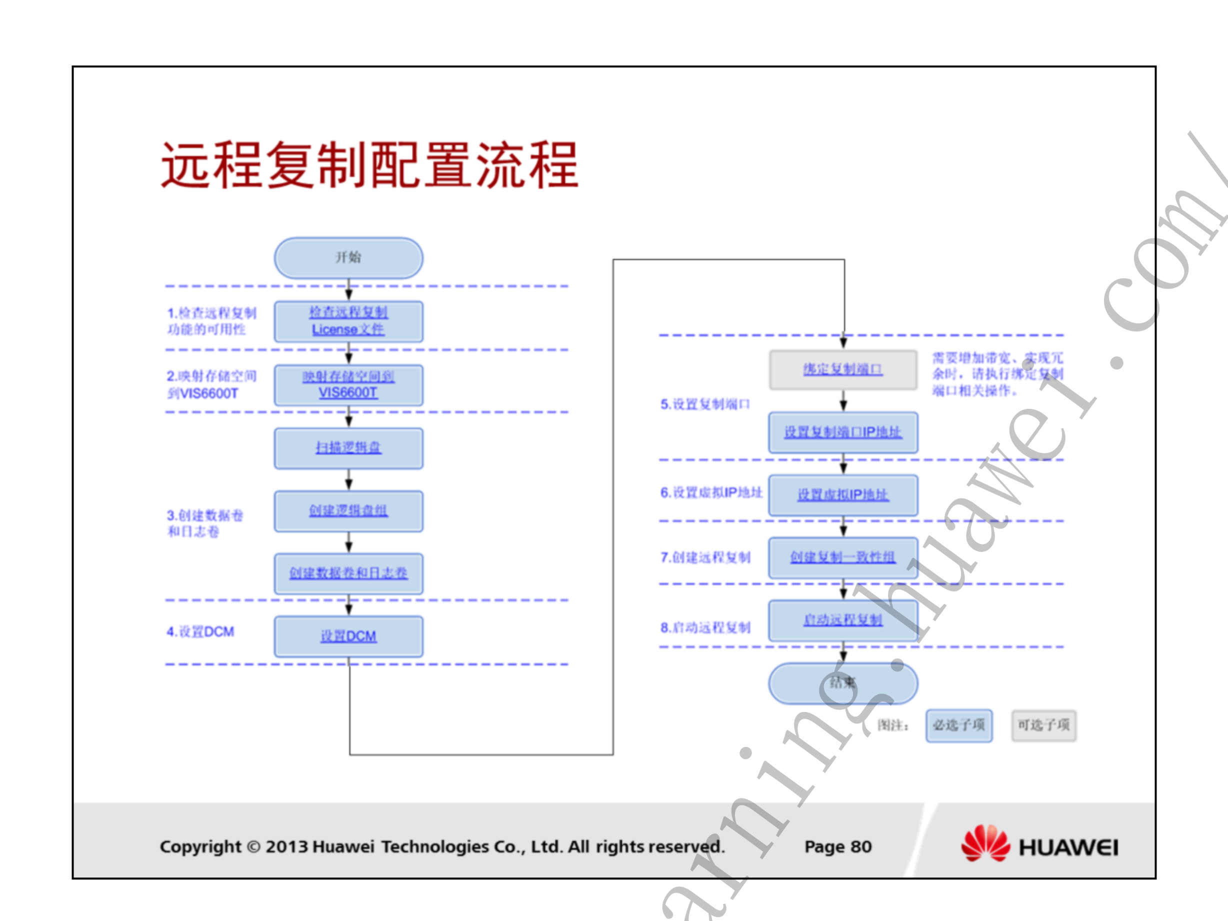
- I/O级复制
  - VIS的复制是基于IO的复制，是粒度最小的复制，并且严格按照生产中心主机下发的IO顺序将IO写到灾备中心。在发生灾难时，可以保证丢失的数据量最小。
- 支持多灾备中心
  - VIS既可以支持多对一的复制，也支持一对多的复制，最大可支持31个灾备中心，即生产中心可以同时向31个灾备中心进行复制，最大程度地保证数据的安全。
- 高扩展性
  - VIS的复制节点数支持在线扩展，可以在不影响现有业务地情况下，对节点数进行扩展以提升VIS的处理能力。

 Page 78





在生产中心和灾备中心分别部署两套阵列，阵列间通过IP网络连接，生产中心和灾备中心都需要配置日志卷和数据卷。


远程复制配置流程
开始
1.检查远程复制功能的可用性
检查远程复制License文件
2.映射存储空间到VIS6600T
映射存储空间到VIS6600T
扫描逻辑盘
3.创建数据卷和日志卷
创建逻辑盘组
创建数据卷和日志卷
4.设置DCM
设置DCM
绑定复制端口
需要增加带宽、实现冗余时，请执行绑定复制端口相关操作。
5.设置复制端口
设置复制端口IP地址
6.设置虚拟IP地址
设置虚拟IP地址
7.创建远程复制
创建复制一致性组
8.启动远程复制
启动远程复制
结束
图注：
必选子项
可选子项
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 80
HUAWEI

远程复制配置参数

- 日志卷
  - 用于存储所有应用程序写入的卷，是复制一致性组的循环写入缓冲区。每次对生产中心的写入都生成两次写入：其中一次写入将写入到日志卷，而另一次写入则写入到数据卷。因此，必须在不同的物理硬盘上配置数据卷和日志卷，以便提高写入性能，且日志卷的大小至少是110MB。
- DCM
  - DCM用于在日志卷溢出时跟踪写入操作，这样可避免对灾备中心上的数据完全重新进行同步。请选择2个逻辑盘用作DCM。
- 虚拟IP地址
  - 对于每个复制卷组，生产中心和灾备中心应该各自有一个唯一的虚拟IP地址用于建立连接。更改生产中心或灾备中心的虚拟IP地址需要暂停复制。



远程复制配置参数-日志卷

- 日志卷
  - 日志卷中包括了若干数据包（默认为8400Byte）和其对应的头信息（默认为512Byte），为确保日志卷的安全，建议在创建日志卷时，使用与创建数据卷不同的硬盘创建日志卷。日志卷不仅需要跟踪生产中心应用数据的写入顺序，还要用于保存异步复制链路断网（包括远程设备下电）期间主机下发的新数据，以便在异步复制链路恢复时，将数据从日志卷中复制到灾备中心。因此日志卷大小的设置可以根据数据卷的数据更新量和容灾网络允许的最大断网时间两个方面进行规划。

 Page 82



- 大多数业务有一个繁忙时间段，其他时间的业务速率不及繁忙时间的10%。同时，由于大多数业务以天为单位，每天业务相差不大，并允许以天为单位的断网时间。
- 日志卷容量大小的计算的两种公式如下:
  - 根据复制一致性组的数据更新量：日志卷大小 > (繁忙业务时间 + 空闲业务时间 / 10) * 峰值写业务速率 *(1 + 512 / 8400)
  - 根据允许断网天数：日志卷大小 > 7 * 复制一致性组每天的数据变化量 * (1 + 512 / 8400)或日志卷大小 > 最大允许断网天数 * 复制一致性组每天的数据变化量 * (1 + 512 / 8400)

远程复制配置参数-DCM

- DCM
  - 用于创建DCM的2个逻辑盘大小相同，且用属于不同的RAID组的逻辑盘。如果有多台存储阵列，推荐用作DCM的2个逻辑盘来源于不同的存储阵列。如果只有一台存储阵列，推荐用作DCM的2个逻辑盘来源于不同的硬盘。
  - 用作DCM的2个逻辑盘的容量大小要大于创建DCM所需的容量。DCM的大小取决于数据卷的大小，DCM默认大小为4KB~256KB，但也可以指定DCM最大为2MB。目前推荐创建DCM使用默认大小，并推荐在一个逻辑盘组中仅创建一个复制一致性组，一个复制一致性组中最多包含128个数据卷，则DCM容量大小 = 128 * 256KB = 32MB。



远程复制配置参数-iSCSI主机端口

- iSCSI主机端口用于建立生产中心和灾备中心之间的复制链路，在生产中心和灾备中心均需配置，规划的项目包括复制端口、复制端口是否需要绑定、复制端口的IP地址等。

Page 84


- 端口： 用于数据复制的iSCSI主机端口，例如P0。
- 是否绑定端口： 为增加数据复制传输带宽，将多个iSCSI主机端口绑定在一起，共同提供数据复制通道。
- IP地址： 设置在iSCSI主机端口上，用于数据复制的IP地址，绑定的端口共同使用同一个IP地址，例如11.11.11.10。

复制状态切换

数据状态	状态说明	是否允许切换
初始同步	复制一致性组在第一次进行同步操作时，数据状态为初始同步中，此时灾备中心的数据是不可用的。	不允许
复制	复制一致性组在进行复制操作时，数据状态为复制状态，此时灾备中心的数据是不可用的。	不允许
溢出	复制一致性组出现故障导致日志卷溢出，数据状态为溢出状态，此时灾备中心的数据与生产中心的数据是不一致性的。如果生产中心无法快速恢复，灾备中心可以接管业务，但存在数据丢失的风险。	允许
重同步	复制一致性组故障恢复后，启用了DCM重同步，数据状态为重同步，此时灾备中心的数据是不可用的。	不允许
停止	复制一致性组在生产中心和灾备中心之间的数据完全一致，数据状态为停止。	允许



当灾难发生时，复制一致性组需要根据生产中心和灾备中心的数据状态判别是否允许进行灾备切换。

VIS6600T产品复制功能应用

- 方案特点
  - 节约带宽
  - 存储整合
  - 不影响主机性能
  - 统一容灾
  - 高可靠性



思考题

1. 存储虚拟化有哪些分类及层次?
2. VIS的安装配置流程是什么?





总结

- 存储虚拟化的概念及其价值
- 存储虚拟化技术实现方式及其优缺点
- 虚拟化存储网关系统产品和应用场景
- 虚拟化存储网关系统产品的软硬件组成和典型组网规划
- 虚拟化存储网关系统产品的基本业务配置和操作
- 虚拟化存储网关系统高级业务

 Page 88





习题

- 判断题
  1. RAID2.0是属于基于主机层的虚拟化技术。（T of F）
- 多选题
  1. 存储虚拟化按实现方式分可以分为？（　　）
     A.带内虚拟化
     B.带外虚拟化
     C.网络虚拟化
     D.块虚拟化



- 习题答案：
  - 判断题：1.F
  - 多选题：1.AB

术语表(1)

- Plex
  - A plex is a logical grouping of subdisks that creates an area of disk space independent of physical disk size or other restrictions. Mirroring is set up by creating multiple data plexes for a single volume. Each data plex in a mirrored volume contains an identical copy of the volume data.
- subdisk
  - A consecutive set of contiguous disk blocks that form a logical disk segment.
  - Subdisks can be associated with plexes to form volumes.
- DCO (data change object)
  - A VxVM object that is used to manage information about the FastResync maps in the DCO volume. Both a DCO object and a DCO volume must be associated with a volume to implement Persistent FastResync on that volume.

 Page 90 

- Plex 是子磁盘的逻辑分组，用于创建不受物理磁盘大小或其他限制约束的磁盘空间区域。可通过为单个卷创建多个数据 plex 来设置镜像。镜像卷中的每个数据 Plex都包含卷数据的相同副本。还可以创建 Plex 来表示连续、条带式和 RAID-5 卷布局以及存储卷日志。
- subdisk（子磁盘） 形成逻辑磁盘区段的一组连续的磁盘块。子磁盘可与 plex 关联以形成卷。
- DCO（数据更改对象） 一个 VxVM 对象，用于管理有关 DCO 卷中快速重同步映射的信息。必须同时将DCO 对象和 DCO 卷与某个卷关联才能在此卷上实现持久性快速重同步。

术语表(2)

- VxVM
  - Veritas Volume Manager by Symantec is a storage management subsystem that allows you to manage physical disks as logical devices called volumes.
- Volume
  - A virtual disk, representing an addressable range of disk blocks used by applications such as file systems or databases. A volume is a collection of from one to 32 plexes.
- DCM (Data Change Map)
  - An object containing a bitmap that can be optionally associated with a data volume on the Primary RVG. The bits represent regions of data that are different between the Primary and the Secondary. The bitmap is used during synchronization and resynchronization.

 Page 91 

- Veritas Volume Manager (VxVM) 是一个存储管理子系统，使用该系统可将物理磁盘作为一种称为“卷”的逻辑设备进行管理。
- Volume（卷） 即虚拟磁盘，表示由文件系统或数据库等应用程序使用的可寻址磁盘块范围。卷是由 1-32 个 plex 组成的集合。
- DCM（Data Change Map，数据更改映射）：一个包含位映射的对象，可选择将该对象与主节点 RVG 中的数据卷关联。位表示主节点和辅助节点之间的不同数据区域。位映射在同步和重新同步过程中使用。

Thank you

www.huawei.com

HC120920004 统一存储与主机连接




HC120920004
统一存储与主机连接
www.huawei.com
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



目标

- 学完本课程后，您将能够:
  - 掌握存储与UNIX主机的连接
  - 掌握UltraPath多路径部署与管理





目录



AIX系统IP-SAN连接常用命令

lsdev -Cc if　检查系统网卡信息

ifconfig –a　检查网卡IP信息

smitty chinet　修改网卡IP地址

lslpp -l *iscsi*　检查iscsi inititator软件是否可用

smitty device　修改initiatorname

编辑/etc/iscsi/targets添加目标器

 Page 3



系统管理界面工具：SMIT-SYSTEM MANAGEMENT INTERFACE TOOL，AIX提供图形用户界面的SMIT管理工具，帮助用户进行外设、终端、备份、安全性、安装及管理等工作，具有菜单驱动，联机帮助、功能扩展、支持MOTIF界面技术、以及分布式平台等特性。

AIX的initiator软件是默认已经在系统中安装好的，如果没有安装请参考AIX的材料进行安装。

修改initiatorname时，输入smitty device光标移动到选择“iscsi”，回车确认后，再选择“iscsi protocal device”,确认后再选择“Change / Show Characteristics of an iSCSI Protocol Device”，然后选择可用的initiator设备，最后在iSCSI Initiator Name条目中输入initiatorname回车确认即可。

vi /etc/iscsi/targets添加如下一行

192.168.1.10 3260 iqn.1995-03.com.quantun:cx0853akr02101-ep1

其中：

192.168.1.10为阵列的iscsi端口ip地址。

3260为iscsi连接的端口。

iqn.1995-03.com.quantun:cx0853akr02101-ep1为阵列端口的targetname，这个字段需要提前获取。

AIX系统FC-SAN连接常用命令

- 检查HBA卡是否被系统正确识别

  # lsdev -Cc adapter | grep fcs

- 查看HBA卡的WWN号

  # lscfg -pvl fcs0



- # lsdev -Cc adapter | grep fcs

  fcs0 Available 02-00 4Gb FC PCI Express Adapter (df1000fe)

  fcs1 Available 03-00 4Gb FC PCI Express Adapter (df1000fe)

- # lscfg -pvl fcs0

  fcs0 U789C.001.DQDH419-P1-C2-T1 4Gb FC PCI Express Adapter (df1000fe)

  ………..

  Device Specific.(Z8)........20000000C985E0FC

  ………..

  PLATFORM SPECIFIC

  Name: fibre-channel

  Model: LPe11000

  Node: fibre-channel@0

  Device Type: fcp

  Physical Location: U789C.001.DQDH419-P1-C2-T1

AIX平台下磁盘管理常用命令

- 发现SAN磁盘资源

  # cfgmgr –v

- 检查磁盘是否正常识别

  # lsdev -Cc disk

- 检查磁盘是否正常识别

  #lspv



- # lsdev -Cc disk

  hdisk0 Available 05-08-00 SAS Disk Drive

  hdisk1 Available 05-08-00 SAS Disk Drive

  hwdisk0 Available 01-00-02 Huawei STOR_V2R2 FC Disk Drive

- # lspv

  hdisk0   00c5e3f45ceaa29b   rootvg   active

  hdisk1   00c5e3f48183d296   rootvg   active

  hwdisk0   none   None

AIX平台下多路径

- 安装多路径软件

  #installp －acd UltraPath-41.01.10T01.powerpc_64.bff all

- 查看UltraPath是否安装成功

  # lslpp -l UPforAIX*

- 重启系统，激活多路径

  # shutdown －Fr

 Page 6



AIX的多路径功能需要另外安装随设备提供的多路径软件UltraPath for AIX。

安装UltraPath for AIX之前请先检查UltraPath for AIX兼容性列表，确保使用与AIX系统兼容的UltraPath版本。

安装UltraPath之前可能会需要对系统进行某些修改，请参考UltraPath安装说明书。



目录



Solaris系统IP-SAN常用命令-IP配置

- 检查网卡IP信息

  # ifconfig –a

- 启用iscsi SAN使用的网卡

  # ifconfig e1000g1 plumb

  # ifconfig e1000g1 inet 172.16.218.1 netmask 255.255.0.0 up

  # ifconfig e1000g1

  # echo "172.16.218.1" >/etc/hostname.e1000g1

- 说明:
  - 如果要永久配置网口IP信息，需要创建/etc/hostname.e1000g1
  - 将IP信息记录在/etc/hostname.e1000g1文件中。

 Page 8



- Solaris for iSCSI SAN的设置过程包含以下几个主题:
  - 在Solaris下设置网卡信息。
  - 在Solaris下检查iSCSI initiator软件。
  - 在Solaris下设置iSCSI initiator信息。
  - 在Solaris下设置iSCSI target信息。
  - 发现iSCSI磁盘设备。

Solaris系统IP-SAN常用命令-iSCSI软件

- 检查iscsi inititator软件是否可用

  # pkginfo | grep -i iscsi

- 查看iscsiadm是否可用

  # /usr/sbin/iscsiadm

 Page 9



iSCSI initiator默认在系统中是已安装软件。如果没有安装，需要安装solairs的操作手册进行安装。

pkginfo | grep -i iscsi

system SUNWiscsir   Sun iSCSI Device Driver (root)

system SUNWiscsitgtr Sun iSCSI Target (Root)

system SUNWiscsitgtu Sun iSCSI Target (Usr)

system SUNWiscsiu   Sun iSCSI Management Utilities (usr)

/usr/sbin/iscsiadm

Usage: iscsiadm -?,-V,--help

Usage: iscsiadm add [-?] <OBJECT> [-?] [<OPERAND>]

Usage: iscsiadm list [-?] <OBJECT> [-?] [<OPERAND>]

Usage: iscsiadm modify [-?] <OBJECT> [-?] [<OPERAND>]

Usage: iscsiadm remove [-?] <OBJECT> [-?] [<OPERAND>]

For more information, please see iscsiadm(1M)

Solaris系统IP-SAN常用命令-iSCSI配置

- 检查iscsi inititator name

  # iscsiadm list initiator-node

- 设置iscsi inititator name

  # iscsiadm modify initiator-node -N iqn.1986-03.com.sun:01

  # iscsiadm modify initiator-node -A sft2002.hselab.com

  # iscsiadm list initiator-node

- 设置iscsi target address

  # iscsiadm add discovery-address 172.16.1.101

  # iscsiadm list discovery-address

  # iscsiadm modify discovery –t enable

  # iscsiadm list discovery

Page 10



iscsiadm list initiator-node　检查iscsi inititator name

Initiator node name: iqn.1986-03.com.sun:01:002128149490.4a0ce9f0

Initiator node alias: -

iscsiadm modify initiator-node -N iqn.1986-03.com.sun:01　　设置iscsi inititator name

iscsiadm modify initiator-node -A sft2002.hselab.com

iscsiadm list initiator-node

Initiator node name: iqn.1986-03.com.sun:01

Initiator node alias: sft2002.hselab.com

iscsiadm add discovery-address 172.16.1.101　　增加iscsi target name

iscsiadm list discovery-address

Discovery Address: 172.16.1.101:3260

iscsiadm modify discovery –t enable　设置iscsi target发现方式

iscsiadm list discovery

Discovery:

Static: disabled

Send Targets: enabled

iSNS: disabled

Solaris系统FC-SAN连接常用命令

cfgadm –a 查看硬件设备

luxadm -e port 检查HBA连接是否正常

fcinfo hba-port 查看HBA卡WWN号

修改/kernel/drv/sd.conf文件，设置多LUN的识别（可选）

update_drv -f sd

 Page 11 

- 若有LUN 0 以外的LUN 从阵列映射给solaris服务器。则需要修改 /kernel/drv/sd.conf文件，将LUN0以外的磁盘写入将文件

```
# vi /kernel/drv/sd.conf
name="sd" class="scsi" class_prop="atapi"  target=0 lun=0;
name="sd" class="scsi" class_prop="atapi" target=0 lun=1;
```

- # fcinfo hba-port

```
HBA Port WWN: 2100001b32870331
OS Device Name: /dev/cfg/c4
Manufacturer: QLogic Corp.
Model: 375-3355-02
Firmware Version: 4.04.01
FCode/BIOS Version:  BIOS: 1.24; fcode: 1.24; EFI: 1.8;
Serial Number: 0402G00-0850667606
Driver Name: qlc
Driver Version: 20080617-2.29
Type: unknown
```

Solaris系统FC-SAN连接常用命令

cfgadm －a 查看硬件设备

luxadm -e port 检查HBA连接是否正常

fcinfo hba-port 查看HBA卡WWN号

修改/kernel/drv/sd.conf文件，设置多LUN的识别（可选）

update_drv -f sd



State: offline

Supported Speeds: 1Gb 2Gb 4Gb

Current Speed: not established

Node WWN: 2000001b32870331



- cfgadm

  “fc-private”代表两块HBA卡（这里是直连组网，没有交换机参与，如果采用有交换机的组网方式，该类型应该会显示为fc-fabric），且其连接状态均为“connected”，驱动程序状态均为“configured”，说明驱动已经加载成功。

- format

  Searching for disks...done

  AVAILABLE DISK SELECTIONS:

  0. c0t0d0 <SUN72G cyl 14087 alt 2 hd 24 sec 424>
  /pci@780/pci@0/pci@9/scsi@0/sd@0,0

  1. c0t1d0 <SUN72G cyl 14087 alt 2 hd 24 sec 424>
  /pci@780/pci@0/pci@9/scsi@0/sd@1,0

  2. c3t200400A0B832927Bd0 <ENGENIO-INF-01-00-0619 cyl 5118 alt 2 hd 64 sec 64>

  /pci@7c0/pci@0/pci@8/SUNW,qlc@0/fp@0,0/ssd@w200400a0b832927b,0

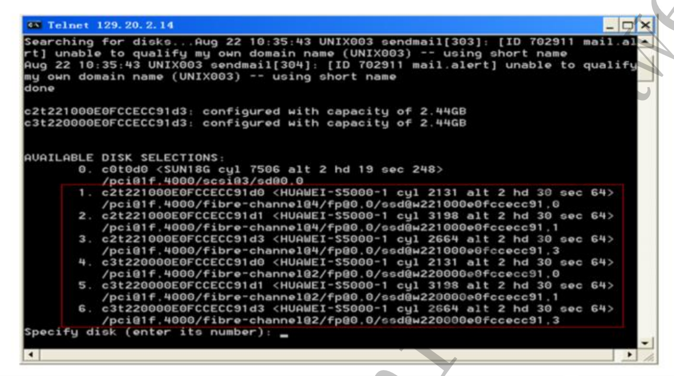
  3. c3t200400A0B832927Bd1 <ENGENIO-INF-01-00-0619 cyl 5118 alt 2 hd 64 sec 64>

  /pci@7c0/pci@0/pci@8/SUNW,qlc@0/fp@0,0/ssd@w200400a0b832927b,1

  Specify disk (enter its number):

Solaris平台下多路径

- Solaris平台的多路径实现采用Solaris自带的STMS功能。
- STMS（StorEdge Traffic Manager Software）由SUN公司开发的随Solaris系统一起发布的一款多路径软件。对于Solaris 8/9，STMS以安装包的形式发布，对于Solaris 10/11，STMS被集成在系统中，随系统一起发布。STMS功能可以支持多路径、Failover、Failback（可选）功能。对于A/A模式的存储控制器，STMS不能实现Failback功能.



 Page 14



使用format命令发现磁盘后，输入磁盘前的LUN号，再输入inquiry，需要记录Vendor和Product字段的内容：

format> inquiry

Vendor: HUAWEI

Product: S5000

修改/kernel/drv/scsi_vhci.conf文件。在/kernel/drv/scsi_vhci.conf中将存储阵列的信息加入，示例如下：

#vi /kernel/drv/scsi_vhci.conf

device-type-scsi-options-list =

"HUAWEI S5000", "symmetric-option";

symmetric-option = 0x1000000;

其中"HUAWEI"为存储阵列的Vendor，该字段必须为八个字符，对于未满八个字符的，需要用空格补齐。"S5000"为存储阵列的Product。

输入以下命令激活光纤卡的多路径功能。Solaris将会重新启动：

stmsboot -D fp –e

重启完成后，使用format命令查看磁盘，以/scsi_vhci/为首的磁盘设备表示为多路径下的虚拟磁盘。

stmsboot –L命令可以查看多路径信息，mpathadm list lu可以查看多路径链路。



目录



HP-UX系统IP-SAN常用命令-网络

ifconfig lan0　　检查网口IP信息

- 通过sam修改网口的IP地址:
  - 在HP-UX服务器上，以超级用户身份登录。
  - 在hpterm或xterm窗口中输入sam。
  - 选择“Networking and Communications”。
  - 选择“Network Interface Cards ”。
  - 选择要激活的网卡。
  - 在SAM菜单栏中选择“Action”→“modify”。
  - 网口配置窗口输入IP信息。
  - 选择“OK”完成网口IP信息设置。



HP-UX系统IP-SAN常用命令-initiator

- 查看软件仓库的iscsi inititator软件

  # swlist -d @ /opt/iSCSI-00_B.11.23.03f_HP-UX_B.11.23_IA_PA.depot

- 检查iscsi inititator软件是否可用

  # swinstall -x autoreboot=true －s /tmp/iSCSI-00_B.11.23.03f_HP-UX_B.11.23_IA_PA.depot iSCSI-00

- 安装iscsi initiator软件

  # PATH=$PATH:/opt/iscsi/bin　为initiator软件增加执行路径

  # iscsiutil　－l 查看initiatorname

  # iscsiutil -i -N iqn.1986-03.com.hp:rx2660-1　修改initiatorname

  # iscsiutil -a -I 172.16.1.101　添加target地址

  # iscsiutil　－pD　查看initiator与目标器的连接

 Page 17



- 安装iSCSI软件过程中需要重新启动。
- # iscsiutil -pD

Discovery Target Information

Target # 1

IP Address : 172.16.1.101

iSCSI TCP Port : 3260

iSCSI Portal Group Tag : 1

User Configured:

Authenticaton Method :

CHAP Method : CHAP_UNI

Initiator CHAP Name :

CHAP Secret :

Header Digest : None,CRC32C (default)

Data Digest : None,CRC32C (default)

HP-UX系统FC-SAN常用命令

ioscan -fnC fc　检查FC HBA卡是否被HP-UX系统识别

fcmsutil /dev/fcd0　查看FC HBA卡的WWN号



- # ioscan -fnC fc

Class　I H/W Path　Driver　S/W State　H/W Type　Description

==

fc　0 0/2/1/0/4/0 fcd　CLAIMED　INTERFACE　HP

AD193-60001 PCI/PCI-X Fibre Channel 1-port 4Gb FC/1-port 1000B-T Combo Adapter (FC Port 1)

/dev/fcd0

- # fcmsutil /dev/fcd0

……….

Link Speed = 4Gb

Local N_Port_id is = 0x000001

Previous N_Port_id is = None

Local Loop_id is = 125

N_Port Node World Wide Name = 0x50014380017abd31

N_Port Port World Wide Name = 0x50014380017abd30

Switch Port World Wide Name = N/A

HP-UX系统FC-SAN常用命令

ioscan -fnC fc　检查FC HBA卡是否被HP-UX系统识别

fcmsutil /dev/fcd0　查看FC HBA卡的WWN号



Switch Node World Wide Name = N/A

Driver state = ONLINE

Hardware Path is = 0/2/1/0/4/0

Maximum Frame Size = 2048

....

HP-UX平台下磁盘管理常用命令

ioscan -H 255　发现目标设备分配的iSCSI资源

insf -H 255　更新ip-san系统磁盘资源

insf -e -C disk　发现fc-san的硬盘

ioscan -funC disk　检查硬盘资源

diskinfo /dev/rdsk/c5t0d0　查看硬盘状态信息

 Page 20



- # insf -e -C disk

insf: Installing special files for sdisk instance 3 address 0/0/2/1.0.16.0.0

insf: Installing special files for sdisk instance 1 address 0/1/1/0.0.0.0.0

insf: Installing special files for sdisk instance 2 address 0/1/1/0.0.0.1.0

insf: Installing special files for sdisk instance 5 address 0/2/1/0/4/0.8.0.2.0.0.0

- # ioscan -funC disk

Class　I H/W Path　Driver　S/W State　H/W Type　Description

==

Disk　5 0/2/1/0/4/0.8.0.2.0.0.0 sdisk　CLAIMED　DEVICE　ENGENIO INF-01-00

/dev/dsk/c5t0d0 /dev/rdsk/c5t0d0

- # diskinfo /dev/rdsk/c5t0d0

SCSI describe of /dev/rdsk/c5t0d0:

vendor: ENGENIO

product id: INF-01-00

type: direct access

size: 20971520 Kbytes

bytes per sector: 512

HP-UX系统多路径

- HP-UX的多路径实现由系统自带的功能实现
  - LVM PVLINK
  - 本地多路径(Native Multipathing)

 Page 21



自 HP-UX 11i v3 起，海量存储堆栈支持本地多路径，而不使用 LVM pvlink。本地多路径比 LVM提供了更多的负载均衡算法和路径管理选项。HP 建议使用本地多路径而不是 LVM 的备用链路来管理多路径设备。为实现向后兼容，可以使用现有的 pvlink。但必须使用物理卷的 Legacy 设备专用文 件，并使用scsimgr 命令禁用这些 Legacy 设备专用文件的本地多路径功能。本地多路径为即开即用，不需要做任何配置。

- pvlink配置

例如，如果磁盘有两条路径，要将一条用作主链路，另一条用作备用链路，请输入以下命令： # vgcreate /dev/vg01 /dev/dsk/c3t0d0 /dev/dsk/c5t0d0

- 配置 LVM
  - 要添加到物理卷的备用链路（该物理卷已经是卷组的一部分），请使用 vgextend 指定指向该物理卷的新链路。例如，如果 /dev/dsk/c2t0d0 已经是卷组的一部分，但还希望再添加另一个指向物理卷的连接，请输入以下命令：

    # vgextend /dev/vg02 /dev/dsk/c4t0d0

  - 如果主链路发生故障，LVM 将自动从主控制器切换到备用控制器。但是，也可以随时使用pvchange 命令要求 LVM 切换到另一个控制器。例如：

    # pvchange -s /dev/dsk/c2t1d0

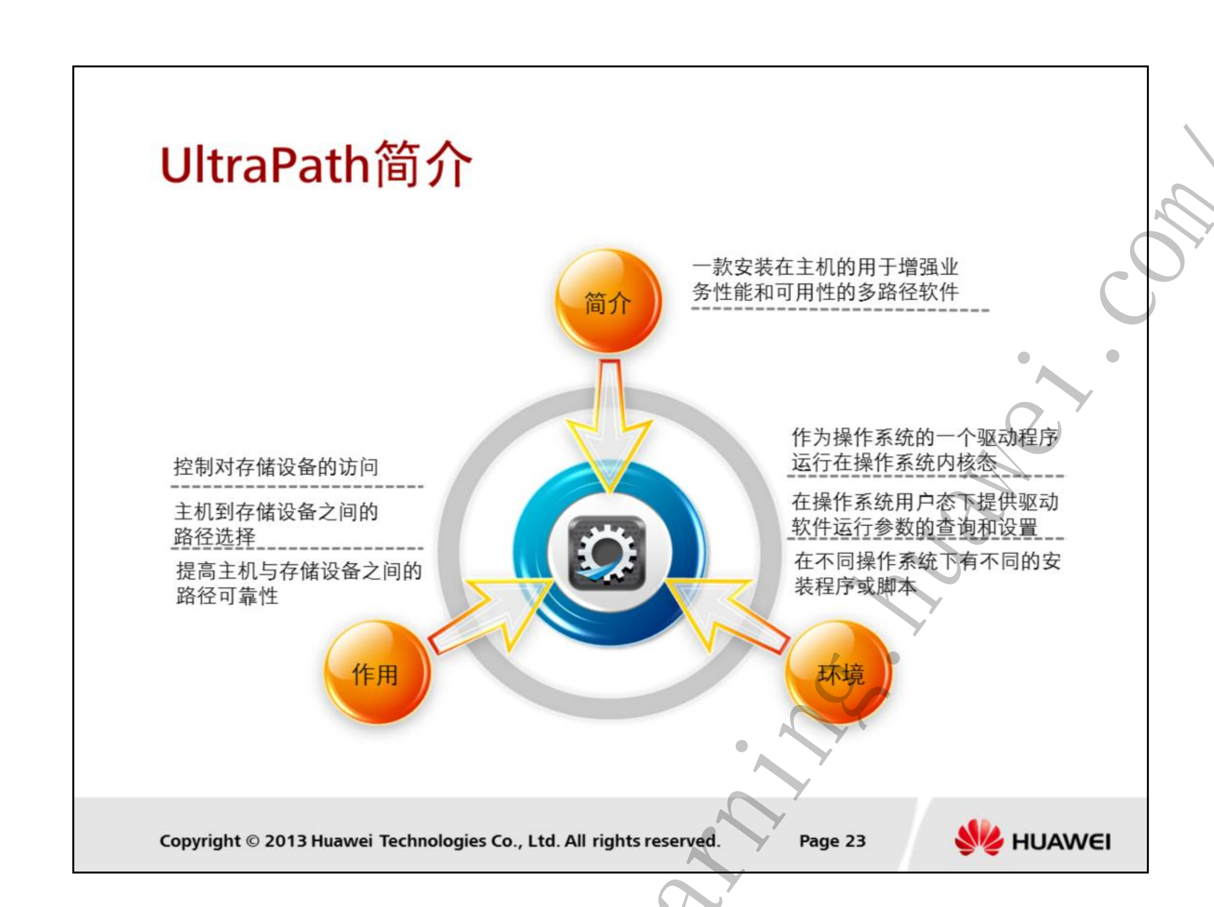
  - 主链路恢复后，LVM 将自动从备用控制器切换回原来的控制器，除非已使用 pvchange 指示它不进行切换，如下所示：

    # pvchange -S n /dev/dsk/c2t2d0

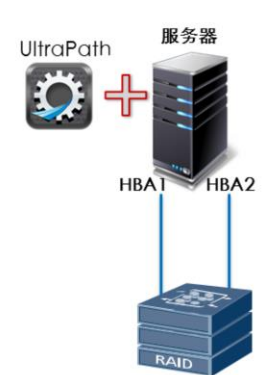


目录




UltraPath简介
简介
一款安装在主机的用于增强业务性能和可用性的多路径软件
控制对存储设备的访问
主机到存储设备之间的路径选择
提高主机与存储设备之间的路径可靠性
作用
作为操作系统的一个驱动程序运行在操作系统内核态
在操作系统用户态下提供驱动软件运行参数的查询和设置
在不同操作系统下有不同的安装程序或脚本
环境
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 23
HUAWEI

实现原理



没有多路径

描述

在没有多路径的情况下，通过操作系统看到设备上的LUN A是两份（分别用LUNA和LUNA`表示）。

风险

通过服务器同时对LUNA和LUNA`写入数据的时候，在设备上会产生冲突。

使用多路径

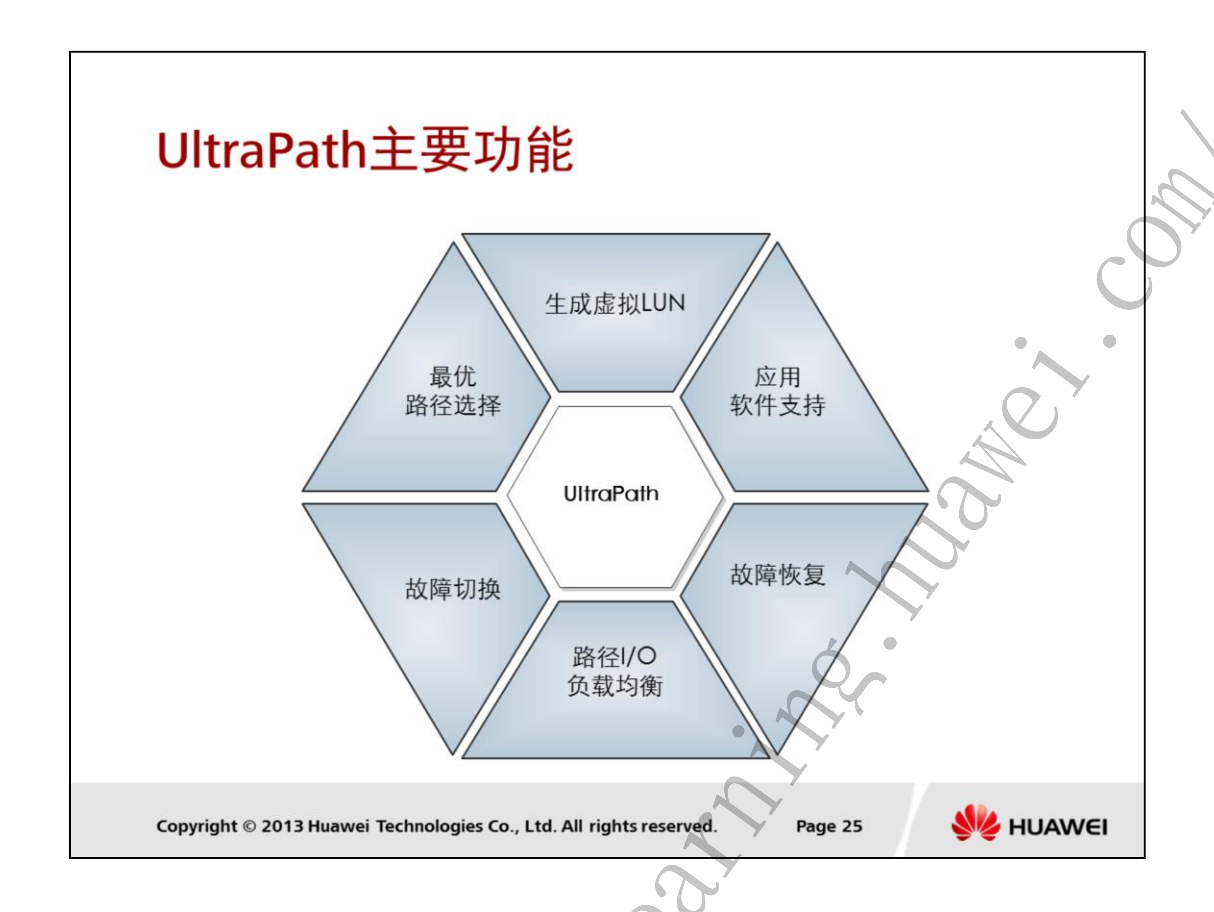
描述

在装有多路径的情况下，两个相同的LUN被虚拟成一个新的设备，新的设备再去对应多条链路，应用程序则使用虚拟盘VDisk0。

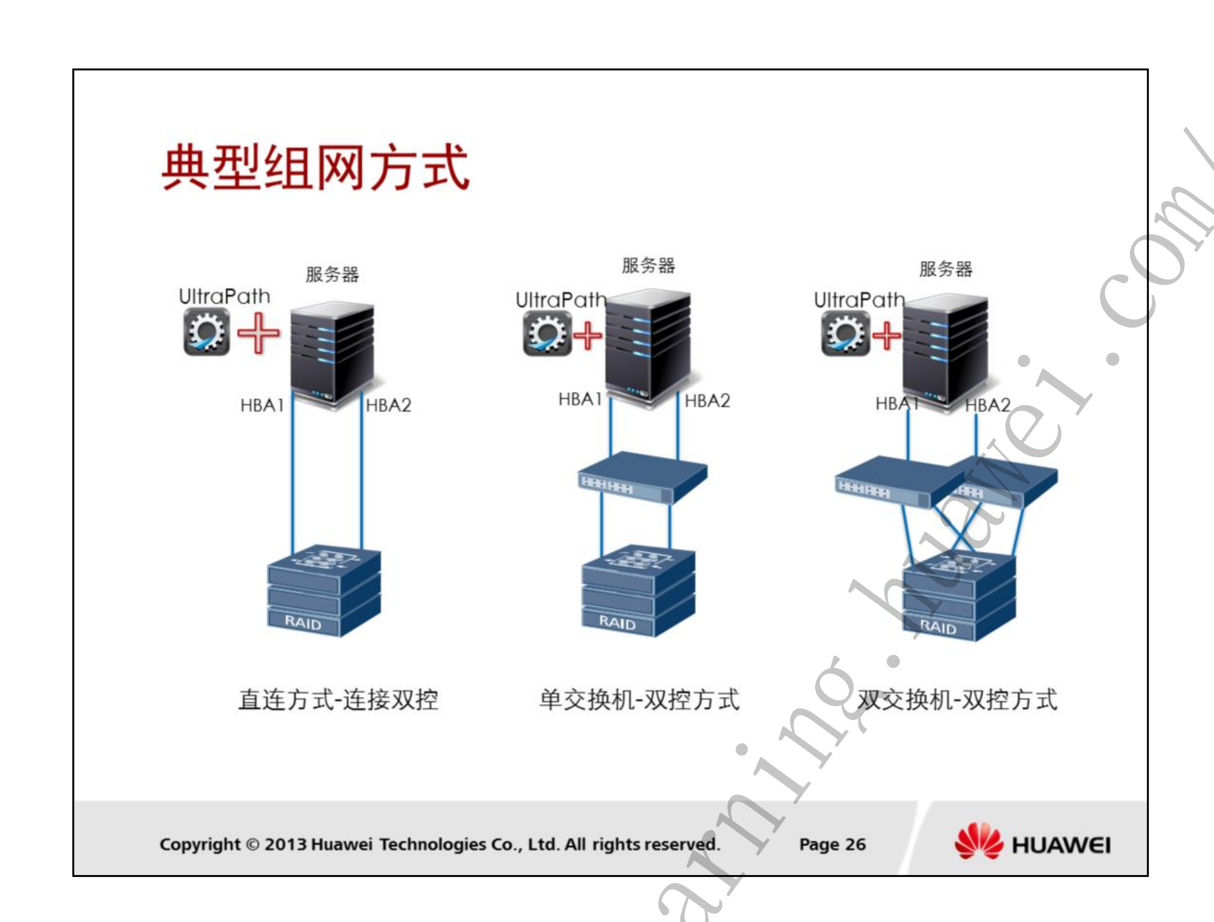


优势

对虚拟LUN进行操作，从而避免了冲突。



- 生成虚拟LUN
  - 对上层用户透明，屏蔽物理LUN，读写操作通过虚拟LUN进行。
- 应用软件支持
  - 支持业界主流集群软件：MSS MSCS、VCS、HACMP、Oracle RAC等集群软件。
  - 支持主流数据库：Oracle、DB2、MySQL、Sybase、Informix等。
- 故障恢复
  - 故障倒换之前的业务链路恢复后，业务恢复（FailBack）随之发生，用户无需介入，自动完成，且业务不中断。
- 路径I/O负载均衡
  - 自动选择多条路径进行下发，提高IO性能。
  - 根据路径繁忙程度进行业务路径选择。
- 故障切换
  - 业务链路发生故障的时候，故障切换（ FailOver ）随之发生，实现业务不中断。
- 最优路径选择
  - 通过LUN归属控制器上的路径进行操作，能获取最佳的性能。


典型组网方式
服务器
UltraPath
HBA1
HBA2
RAID
直连方式-连接双控
单交换机-双控方式
双交换机-双控方式
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 26
HUAWEI



目录

HUAWEI



UltraPath for Solaris安装步骤

下表列出了UltraPath for Solaris的安装步骤：

	安装步骤
安装前准备	1.使用具有Solaris操作系统超级用户（root）权限的账号。
	2.确保操作系统是sparc架构下的Solaris 10 Update8, Solaris11 Update 1。
	3.保证服务器上/opt目录至少有200M剩余空间。
	4.只支持FC组网。
开始安装	1.拷贝安装包到目标机。
	2.执行安装。
安装后检查	1.重启主机使多路径生效。
	2.通过upadm检查安装的多路径版本。



UltraPath for Solaris安装前准备

- 使用具有Solaris操作系统超级用户（root）权限的账号登录访问系统：
- 通过命令cat /etc/release查看系统版本，确保操作系统是sparc架构下的Solaris 10 Update8。
- 通过命令df -h /opt/查看存储空间，保证服务器上/opt目录至少有200M剩余空间。

Page 29



使用具有Solaris操作系统超级用户（root）权限的账号登录访问系统：

1、本地访问

[root@Solaris10:/]# who

root   console   Jun 11 11:20

2、远程访问

[root@Solaris10:/]# who

root   pts/1   Jun 11 11:28   (129.42.2.2)

确保操作系统是sparc架构下的Solaris 10 Update8。

[root@Solaris10:/]# cat /etc/release

Solaris 10 10/09 s10s_u8wos_08a SPARC

Copyright 2009 Sun Microsystems, Inc. All Rights Reserved.

Use is subject to license terms.

Assembled 16 September 2009

保证服务器上/opt目录至少有200M剩余空间。

[root@Solaris10:/]# df -h /opt/

Filesystem   size   used   avail capacity Mounted on

/dev/dsk/c1t0d0s0   85G   36G   48G   43%   /

安装UltraPath for Solaris

- 拷贝安装包到目标机；
- 执行安装：切换到安装包目录，执行install.sh，开始安装。
  - 如果检测到自动的STMS多路径管理了OceanStor存储系统，会提示是否关闭STMS对OceanStor存储系统的管理，输入”y”，继续

    Huawei arrays are managed by the STMS.

    Are you sure to cancel the management of Huawei arrays under the STMS to avoid the conflict with the management of Huawei arrays under the UltraPath?(y or n):

 Page 30




```
-bash-3.00# sh install.sh

Checking free space of /opt ...

Transferring <UltraPath> package instance

The following packages are available:

  1  UltraPath     UltraPath for Solaris

 (sparc) 5.01.037

Select package(s) you wish to process (or 'all' to process

all packages). (default: all) [?,??,q]:
```

UltraPath for Solaris安装后检查

- 安装完成后需要重启主机后多路径才能生效

----

* UltraPath Installation:

* Installation is successful.

* Please REBOOT the host to make the installation

* effective!

Installation of <UltraPath> was successful.

- 通过upadm检查安装的多路径版本

[root@chenwei:/export]# upadm

UltraPath CLI #0 >show version

Software Version : 5.01.037

Driver Version : 5.01.037



1.安装完成后需要重启主机后多路径才能生效

----

* UltraPath Installation:

* Installation is successful.

* Please REBOOT the host to make the installation

* effective!

Installation of <UltraPath> was successful.

2.通过upadm检查安装的多路径版本

[root@chenwei:/export]# upadm

UltraPath CLI #0 >show version

Software Version : 5.01.037

Driver Version : 5.01.037

UltraPath for AIX安装前准备-识别硬盘

- 未安装多路径时，区分存储上报的磁盘与本地磁盘。
  - 通过检查命令lscfg -vpl hdiskX的输出来判断，如下所示：

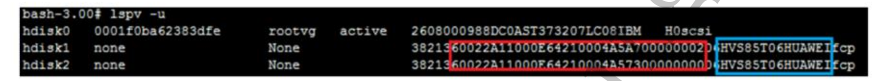
```
bash-3.00# lscfg -vpl hdisk3
  hdisk3           U787F.001.DPM0HW7-P1-C1-T1-W200A0022A10E6421-L1000000000000  Other FC SCSI Disk Drive

        Manufacturer................HUAWEI
        Machine Type and Model......HVS85T
        ROS Level and ID............33313031
        Serial Number...............
        Device Specific.(Z0)........000004323F081002
        Device Specific.(Z1)........


  PLATFORM SPECIFIC

  Name:  disk
    Node:  disk
    Device Type:  block
```

  - 高版本系统中，可以通过检查命令lspv -u的输出来判断，如下所示，其中红框中为磁盘的WWN，蓝框为存储型号及厂商：





UltraPath for AIX安装前准备

- 由于安装完毕后需重启AIX主机，故需要先停止业务。
- 保证应用服务器上“/opt”目录至少有200M剩余空间以及“/usr”目录至少有10M剩余空间。
- 查询应用服务器的AIX版本及操作系统位数，需保证AIX操作系统是5300-03及以上、6100-01及以上或7100-01及以上，并且操作系统的位数为64位。
- 使用命令lsdev -Cc adapter | grep fc，确定应用服务器上HBA卡的状态是“Available”。

 Page 33 

- 安装完毕后需要重启AIX主机，故需要停止业务。
- 保证应用服务器上“/opt”目录至少有200M剩余空间以及“/usr”目录至少有10M剩余空间。

```
# df -m
Filesystem MB blocks Free %Used Iused %Iused Mounted on
/dev/hd2 5248.00 685.33 87% 61338 28% /usr
/dev/hd10opt 2048.00 1852.50 10% 2333 1% /opt
```

- 查询应用服务器的AIX版本及操作系统位数，需保证AIX操作系统是5300-03及以上、6100-01及以上或7100-01及以上，并且操作系统的位数为64位。

```
# oslevel –r 5300-06
# bootinfo –y 64
```

- 确定应用服务器上HBA卡的状态是“Available”。

```
#lsdev -Cc adapter | grep fc
fcs0 Defined 14-08 FC Adapter
fcs1 Defined 1A-08 FC Adapter
fcs2 Available 1A-08 FC Adapter
fcs3 Available 1D-08 FC Adapter
```

说明

fcs0~fcs3表示AIX主机上的4张HBA卡，实际情况HBA卡的数量可能不同。

Defined表示卡不可用，Available表示卡可用。

UltraPath for AIX安装前准备

- 如果在安装UltraPath for AIX软件前，华为存储系统已添加映射LUN给AIX主机且已上报到AIX操作系统，则必须执行rmdev -dl命令将物理磁盘删除（使用lscfg -vpl hdiskXX确定是否华为磁盘）。
- 使用命令lscfg -vpl fcs0，确定应用服务器上HBA卡的型号。保证HBA卡是AIX应用服务器自带的。

 Page 34 

- 如果在安装UltraPath for AIX软件前，华为存储系统已添加映射LUN给AIX主机且已上报到AIX操作系统，则必须执行rmdev -dl命令将物理磁盘删除（使用lscfg -vpl hdiskXX确定是否华为磁盘）。

  # rmdev -dl hdisk2

  hdisk2 deleted

- 使用命令lscfg -vpl fcs0，确定应用服务器上HBA卡的型号。保证HBA卡是AIX应用服务器自带的。

lscfg -vpl fcs0

fcs0 U787B.001.DNWF39F-P1-C3-T1 FC Adapter

Part Number.................03N5014

EC Level....................A

Serial Number...............1F84637031

Manufacturer................001F

Feature Code/Marketing ID...280D

FRU Number.................. 03N5014

…………..

Name: fibre-channel

Model: LP11000

Node: fibre-channel@1

UltraPath for AIX安装前准备

- 如果在安装UltraPath for AIX软件前，华为存储系统已添加映射LUN给AIX主机且已上报到AIX操作系统，则必须执行rmdev -dl命令将物理磁盘删除（使用lscfg -vpl hdiskXX确定是否华为磁盘）。
- 使用命令lscfg -vpl fcs0，确定应用服务器上HBA卡的型号。保证HBA卡是AIX应用服务器自带的。

 Page 35



Device Type: fcp

Physical Location: U787B.001.DNWF39F-P1-C3-T1

其表示该HBA卡的型号为LP11000，对应的IBM标识Feature Code/Marketing ID为280D（在AIX高版本系统中该字段被修改为“Customer Card ID Number”）。

安装UltraPath for AIX

- 将UltraPath for AIX软件包上传到AIX主机的任意目录（一般使用tmp目录），建议新建一个目录存放UltraPath for AIX软件安装包中所有文件，注意：不要改动软件包的目录结构，软件包目录不能含有空格！
- 进入上传到AIX系统的软件包目录，执行tar -xvf *install.tar命令，把tar文件解压为“install.sh”*
- 执行sh *install.sh命令进行安装。*



执行sh *install.sh命令*，开始安装，该脚本首先进行环境检查，例如当存在冗余物理磁盘时，会提示选择继续或者退出安装：

WARNING :The array disk exist in system,please delete it first.

If not,virtual lun may not been made successful.

1 warning exists during installing the UltraPath software,are you sure continue(Y/y)or(N/n)? : -->

当环境检查通过时，则安装结束时系统有UltraPath安装完成的提示信息。

安装UltraPath for AIX

- 执行chdev -l Name -a fc_err_recov=xxx和chdev -l Name -a dyntrk=xxx命令，配置fscsi设备属性，修改属性前需要先将fscsi上的子设备设置为defined状态；
- 执行shutdown -Fr命令，重启系统；
- UltraPath的初始化配置在安装前已经包含在安装包中，安装完成后即会采用默认配置进行工作。



- 首先执行*lsdev -Cc driver | grep fscsi*列出当前的fscsi设备：

```
-bash-3.00# lsdev -Cc driver | grep fscsi
fscsi0 Available 06-08-01 FC SCSI I/O Controller Protocol Device
fscsi1 Available 0B-08-01 FC SCSI I/O Controller Protocol Device
```

- 然后执行lsdev -p fscsiX，分别查询各fscsi设备是否有子设备：

```
-bash-3.00# lsdev -p fscsi0
sfwcomm0 Available 06-08-01-FF Fibre Channel Storage Framework Comm
```

- 根据查询的结果，执行rmdev -l subDevice将子设备设置为defined状态：

```
-bash-3.00# rmdev -l sfwcomm0
sfwcomm0 Defined
```

- 最后修改HBA卡属性：

DAS直连：dyntrk参数：no；fc_err_recov参数：delayed_fail

SAN交换机：dyntrk参数：yes；fc_err_recov参数：fast_fail

- 执行shutdown -Fr命令，重启系统。
- UltraPath的初始化配置在安装前已经包含在安装包中，安装完成后即会采用默认配置。

安装UltraPath for AIX -重启后检查

- 执行upadm show version命令检查已安装版本是否与目标版本一致
- 执行lsdev -Cc disk检查虚拟磁盘数目是否与映射LUN数目一致，若数目不一致，则使用rmdev -dl hdiskXX 命令删除所有的华为磁盘（使用lscfg -vpl hdiskXX确定），再执行cfgmgr命令。
- 执行lspath检查路径数目是否与实际组网一致。
- 执行upadm show lun命令确定各虚拟磁盘与LUN的对应关系。

 Page 38



- 执行upadm show version命令检查已安装版本是否与目标版本一致

upadm show version

Software Version : 6.01.008

Driver Version : 6.01.008

- 执行lsdev -Cc disk检查虚拟磁盘数目是否与映射LUN数目一致，若数目不一致，则使用rmdev -dl hdiskXX 命令删除所有的华为磁盘（使用lscfg -vpl hdiskXX确定），再执行cfgmgr命令。

lsdev -Cc disk

hdisk0 Available 1Z-08-00-8,0 16 Bit LVD SCSI Disk Drive

hdisk1 Available 1Z-08-00-9,0 16 Bit LVD SCSI Disk Drive

hdisk2 Available 1A-08-02 Other FC SCSI Disk Drive

hdisk3 Defined 1A-08-02 Other FC SCSI Disk Drive

hdisk4 Available 07-08-01 Huawei S5500T FC Disk Drive

在如上输出中，需删除hdisk2、hdisk3及hdisk4后执行cfgmgr。

- 执行lspath检查路径数目是否与实际组网一致。
- 执行upadm show lun命令确定各虚拟磁盘与LUN的对应关系。



目录

 Page 39



UltraPath for Windows基本功能配置

- Windows下配置UltraPath有两种方式：
  - upadm命令行配置方式
    - 在Windows系统通过“开始->所有程序->huawei->OceanStor UltraPath->upadm”或直接在Windows　的命令行输入upadm，进入upadm命令行通过输入CLI命令配置UltraPath。
  - UltraPath Console 图形化界面配置方式
    - 在Windows系统通过“开始->所有程序->huawei->OceanStor UltraPath-> UltraPath Console”或
    - 直接在系统桌面点击“UltraPath Console”快捷方式图标进入UltraPath Console控制台配置UltraPath。



UltraPath for Windows基本功能配置

- 下表列出了部分upadm命令行配置常用命令：

命令	功能	参数说明	生效方式
set pathstate={enable | disable} path_id=<ID>	启用/禁用一条物理路径	.	立即
set phypathnormal path_id=<ID>	手动置好一条物理路径	.	立即
set tpgstate={enable | disable} array_id=<ID> tpg_id=<A | B | ID>	启用/禁用一个控制器	.	立即
set workingmode=<ID>	设置阵列的工作模式	0 – 控制器间负载均衡 1 – 控制器内负载均衡	立即
set luntrespass={on | off}	开启/关闭LUN切换功能	.	立即
set failbackdelaytime=<time>	设置延迟Failback时间	0-1200 s	立即
set iosuspensiontime=<time>	设置全网闪断时间	0-600 s	立即
set ioretry=<number> ioretrydelay=<time>	设置IO重试次数与重试延迟时间	重试次数 1-128 次 延迟时间 1-120 s	立即
start pathcheck path_id=<ID1,ID2,...>}	手动发起路径健康度检查	.	立即

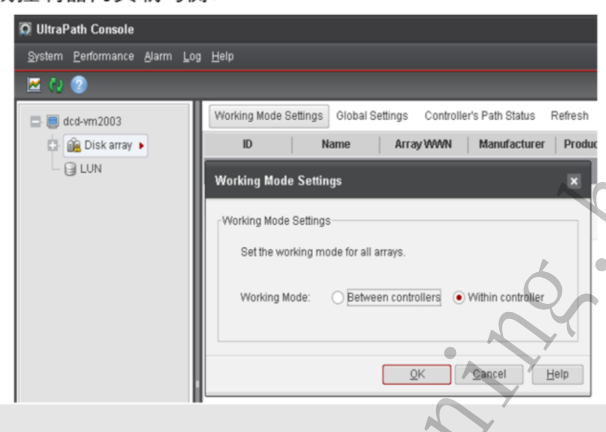
 Page 41



- 在多路径的命令行中输入“？”查看所有命令格式

UltraPath for Windows基本功能配置

- 设置阵列工作模式：
  - 请按照如下顺序 “UltraPath Console->Host->Disk array->Working Mode Settings”进入工作模式设置界面，设置所有阵列的工作模式为控制器间负载均衡或控制器内负载均衡。







 Page 42 HUAWEI

- 在“Working Mode”的右侧单选按钮中，选择Between Controllers，则所有阵列的负载均衡模式被设置为控制器间负载均衡。选择Within Controller，则所有阵列的负载均衡模式被设置为控制器内负载均衡。
- 工作模式默认为“控制器内负载均衡”。

UltraPath for Windows Console配置

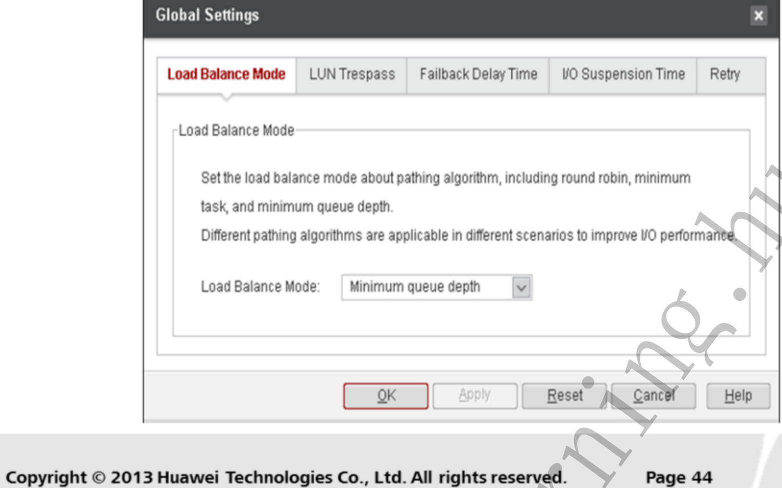
- 设置全局属性：
  - 请按照如下顺序 “UltraPath Console->Host->Disk array->Global Settings”进入全局设置界面，设置所有阵列的负载均衡模式、切LUN功能、Failback延迟时间、IO悬挂时间、重试策略等参数。

 Page 43 HUAWEI

- 在全局设置界面，点选上方的您想设置的属性的Tab头，然后在tab头对应的参数框中设置对应参数。

UltraPath for Windows Console配置

- 设置全局属性(负载均衡模式设置):
  - 通过负载均衡模式设置，可以选择使用轮询、最小队列深度或最小任务选路算法。不同业务场景可以通过设置不同的选路算法，用以提高IO性能。

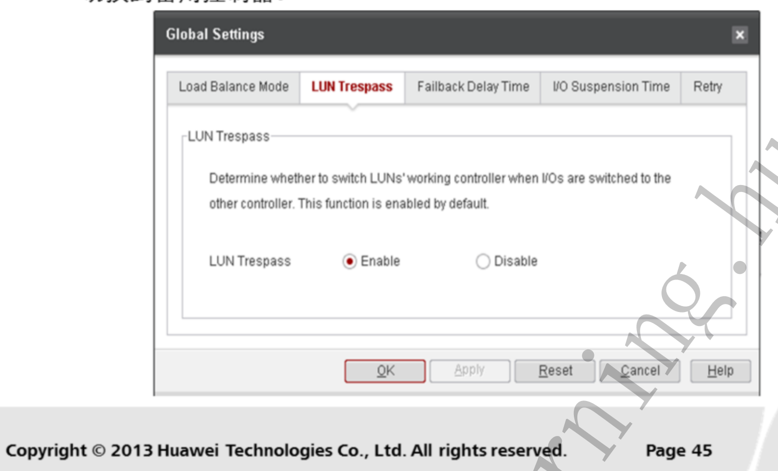


Page 44
HUAWEI

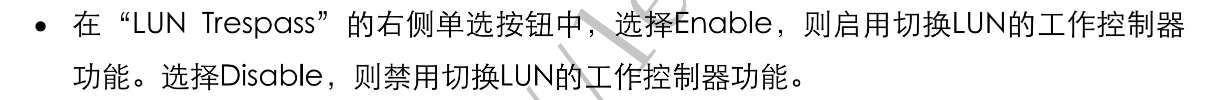
- 在“Load Balance Mode”右侧下拉列表可选择轮询、最小队列深度、最小任务等选路算法，默认值为最小队列深度。
  - 轮询算法：数据在可用路径上轮流下发。
  - 最小队列深度算法：多路径软件会根据当前每条路径上的IO个数，选择所有路径中IO 个数最少的那个路径下发IO 。
  - 最小任务算法：不仅计算IO个数，同时计算每个IO访问的数据量大小，综合选出负载最小的路径下发IO 。

UltraPath for Windows Console配置

- 设置全局属性(切换LUN设置):
  - 通过切换LUN设置，可以启用或禁用切LUN开关。启用后，当多路径优选控制器上的所有路径都发生故障后，把IO 往备用控制器下发，并把LUN 的工作控制器切换到备用控制器。



 Page 45 HUAWEI



- 在“LUN Trespass”的右侧单选按钮中，选择Enable，则启用切换LUN的工作控制器功能。选择Disable，则禁用切换LUN的工作控制器功能。
- 切换LUN设置默认值为“启用”状态。

UltraPath for Windows Console配置

- 设置全局属性(延迟FailBack时间设置):
  - 当多路径检测到优选控制器故障恢复后，可以将I/O 立即在优选控制器继续下发。但可能存在主机和优选控制器间的链路不稳定的情景，则需要设置一定时间间隔，延时将I/O 切回优选控制器，待路径稳定后再将I/O 选择优选控制器下发。

Global Settings

Load Balance Mode | LUN Trespass | Failback Delay Time | I/O Suspension Time | Retry

Failback Delay Time

Set the failback delay time.

If failback delay is disabled, the time changes to 0 seconds automatically.

☑ Enable failback delay

Delay time (s): 600 (1-1200,The value is 600 by default)

OK Apply Reset Cancel Help

 Page 46 HUAWEI

- 勾选 “Enable failback delay” ，并输入Delay time,则启用延迟FailBack时间功能。取消勾选 “Enable failback delay” ，则Delay time被自动设置为0，延迟FailBack时间功能被禁用。此时若多路径检测到优选控制器故障恢复后，立即将 I/O 向优选控制器继续下发。
- 延迟FailBack时间默认为 “启用” 且Delay time 默认为600秒。

UltraPath for Windows Console配置

- 设置全局属性(I/O悬挂时间设置)：
  - 启用I/O悬挂功能并设置I/O悬挂时间后，当主机和阵列之间的路径全部故障，多路径软件可以在设置的时间内将I/O悬挂，而不向上层应用程序报错。

Global Settings

Load Balance Mode | LUN Trespass | Failback Delay Time | I/O Suspension Time | Retry

I/O Suspension Time

Set the I/O suspension time. This function is disabled by default. If I/O suspension is disabled, the time changes to 0 seconds automatically.

When all paths between a host and the array are faulty, the UltraPath can suspend I/Os for a specified period, without reporting errors to upper-layer application programs.

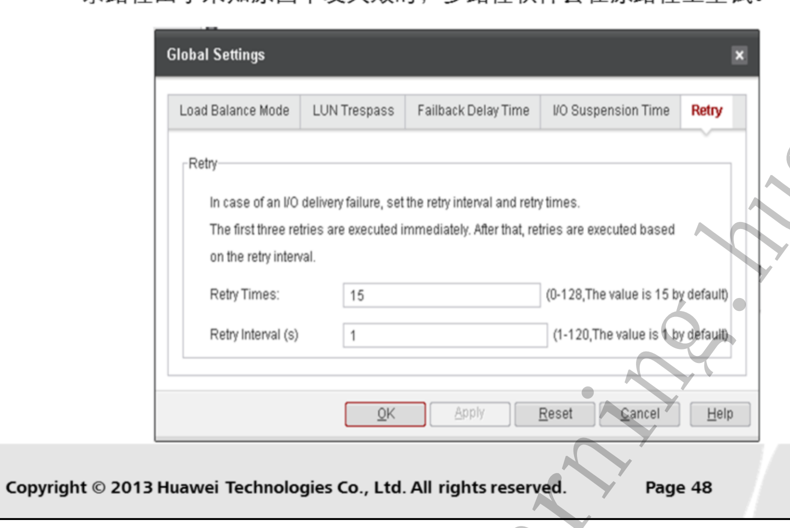
Enable I/O Suspension

I/O Suspension Time (s): 0 (1-600)

OK Apply Reset Cancel Help

 Page 47 HUAWEI

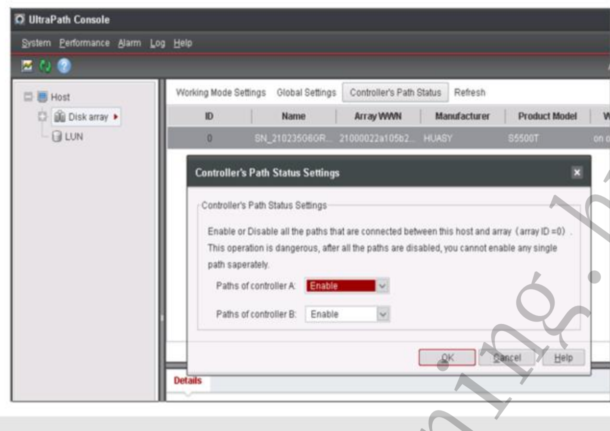
- 勾选 “Enable I/O Suspension” ，并输入I/O Suspension Time,则启用I/O悬挂功能。取消勾选 “Enable I/O Suspension” ，则I/O Suspension Time被自动设置为0，I/O悬挂功能被禁用。此时若主机和阵列之间的路径全部故障，多路径软件立即向上层应用程序报错。
- I/O悬挂功能默认为 “禁用” 状态 。



- 在“Retry Times”的右侧输入框中，输入当设置I/O下发失败后的重试次数。默认为15次。
- 在“Retry Times”的右侧输入框中，输入当设置I/O下发失败后的重试时间间隔。默认为1秒。

UltraPath for Windows Console配置

- 启用/禁用控制器下与主机相连的所有路径：
  - 请按照如下顺序“UltraPath Console->Host->Disk array->Controller’s Path Status”进入控制器路径状态设置界面，启用或禁用控制器下与主机相连的所有路径。

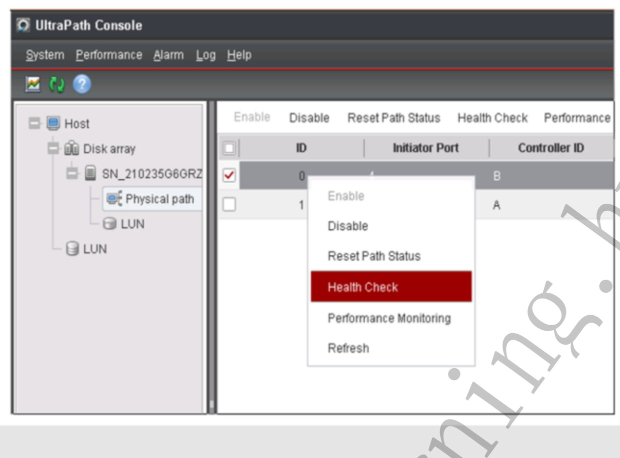


 Page 49 

- 在对应的控制器的右方下拉列表选择启用或禁用，通过该操作可以禁用或启用本主机与阵列上指定控制器所连接的所有物理路径。
- 将控制器路径状态设置为“禁用”状态具有危险性，禁用后，将不能单独启用该控制器下任何一条物理路径。

UltraPath for Windows Console配置

- 启用/禁用物理路径、重置物理路径状态、路径健康度检查
  - 请按照如下顺序 “UltraPath Console->Host->Disk array->Physical path ” 选择目标路径，点击右键进行与物理路径相关的各种操作。



 Page 50 HUAWEI

- 启用/禁用物理路径将主动将路径启用或禁用。
- 重置物理路径状态设置可以手动将路径的状态置为正常状态。如果该路径为禁用状态，则无法强制将其设置为正常状态。
- 路径健康检查功能通过下发测试码流来检测路径的健康状况。当用户发现路径降级之后，采用更换链路的方式进行修复。然后可以通过路径健康检查来确认更换的路径是否正常工作。

UltraPath for Linux基本功能配置

命令	功能	参数说明
upadmin set tpgstate={enable | disable} array_id=<ID> tpg_id=<A | B | ID>	设置阵列控制器状态	
Upadmin set pathstate={enable | disable} path_id=<ID>	设置路径状态	DI：路径ID
upadmin set phypathnormal path_id=<ID>	手动恢复路径正常	ID：路径ID
upadmin set workingmode=[0 | 1]	设置是否开启控制器之间的负载均衡	0: load balancing between controllers 1:load balancing within controller Default is "1"
upadmin set luntrespass={on | off}	设置是否开启切换LUN工作控制器的功能。	.
upadmin set failbackdelaytime=<Time>	设置failback扫描的间隔时间	Time：0~1200s，立即生效，需使用upLinux updateImage进行保存
Upadmin set iosuspensiontime=<Time>	设置io悬挂时间	Time：0~600s，默认是0
Upadmin ioretry=<Number> ioretrydelay=<Time>	设置io重试次数和间隔时间	Number：0~128 Time：1~120s
upadmin set loadbalancemode={round_robin | min-queue-depth | min-task}	设置负载均衡算法。	round-robin：轮询算法 min-queue-depth：最小队列深度算法。 min-task：最小IO负载算法。

 Page 51 

- 如果不修改配置，UltraPath安装完成后将采用默认配置进行工作。
- 更多详细配置信息请参考《OceanStor UltraPath for Linux 用户指南》

UltraPath for Solaris基本功能设置

命令	功能	参数说明	生效方式
upadm set pathstate={enable | disable} path_id=<ID>	启用/禁用一条物理路径	.	立即
upadm set phypathnormal path_id=<ID>	手动置好一条物理路径	.	立即
upadm set tpgstate={enable | disable} array_id=<ID> tpg_id=<A | B | ID>	启用/禁用一个控制器	.	立即
upadm set workingmode=<ID>	设置阵列的工作模式	0 – 控制器间负载均衡 1 – 控制器内负载均衡	立即
upadm set loadbalancemode={round-robin | min-queue-depth | min-task}	设置负载均衡算法	round_robin：轮询算法 min-queue-depth：最小IO队列深度算法。 min-task：最小IO负载算法。	立即
upadm set luntrespass={on | off}	开启/关闭LUN切换功能	.	立即
upadm set failbackdelaytime=<time>	设置延迟Failback时间	0-1200 s	立即
upadm set iosuspensiontime=<time>	设置全网闪断时间	0-600 s	立即
upadm set ioretry=<number> ioretrydelay=<time>	设置IO重试次数与重试延迟时间	重试次数 1-128 次 延迟时间 1-120 s	立即
upadm start pathcheck path_id=<ID1,ID2,...>}	手动发起路径健康度检查	.	立即



- 表列出了UltraPath for Solaris的配置

UltraPath for AIX基本功能设置

命令	功能	举例	说明
upadm help	显示简要帮助信息。	upadm help	显示简要帮助信息。
upadm show version	查询UltraPath for AIX软件版本信息。	upadm show version	查询UltraPath for AIX软件版本信息。
upadm show daemon	查看UltraPath for AIX守护进程运行状态。	upadm show daemon	查看UltraPath for AIX守护进程运行状态。
upadm show option	查看UltraPath for AIX可配置选项。	upadm show option	查看UltraPath for AIX可配置选项。
cfgmgr	扫描硬盘。	cfgmgr	在应用服务器上安装完UltraPath for AIX软件或在存储系统上给应用服务器新添加映射后，请重新扫描硬盘。
lspv	查看硬盘的概要信息。	lspv	在扫描硬盘后，使用lspv命令查看扫描到的硬盘的信息。
upadm show lun [dev=*hdiskxx*]	查看虚拟硬盘的详细信息。	upadm show lun dev=*hdisk5*	查看虚拟硬盘hdisk5的详细信息。
lspath -F	查看硬盘的路径信息。	lspath -F "*name: parent: connection: status: path_id*"	自定义显示硬盘的路径信息格式为从左到右各列显示硬盘的名称、路径的父设备、路径的链接信息、硬盘的状态和路径ID。

 Page 53



- UltraPath的其他初始化配置在安装前已经包含在安装包中，安装完成后即会采用默认配置进行工作。

UltraPath for AIX基本功能设置

命令	功能	举例	说明
lspath	查看路径的优先级属性。	lspath -a *priority* -F *value* -l *hdisk2* -p *fscsi0* -w *43cdcefce0001122,0*	显示虚拟硬盘hdisk2的优先级属性。
chdev -a *Attribute=Value* -l *Name*	配置设备属性。	chdev -a *algorithm=round_robin* -l *hdisk2*	修改虚拟硬盘hdisk2的选路方法为round_robin。
chdev -l *Name* -a *fc_err_recov=fast_fail* chdev -l *Name* -a *dyntrk=yes*	配置HBA卡属性。	chdev -l *fscsi0* -a *fc_err_recov=fast_fail*	配置fscsi0的HBA卡属性为快速错误恢复。
upadm chkconfig	检查UltraPath for AIX相关系统配置。	upadm chkconfig	检查系统的当前配置是否符合UltraPath for AIX的要求。
upadm set vismaxnode=*value*	配置VIS的最大节点个数。	upadm set vismaxnode= *6*	配置VIS的最大节点个数为6。
upadm remove array=*array_type*	删除虚拟硬盘。	upadm remove array=*HW S5300*	删除存储系统型号为HW S5300的虚拟硬盘。
upadm set failover=*on*	开启LUN故障切换功能。	upadm set failover=*off*	关闭LUN故障切换功能。
odmget -q *name = xxxx* CuDv	获取硬盘类型信息。	smit hacmp	进入ReserveCheck和ReserveBreak功能配置界面。

 Page 54

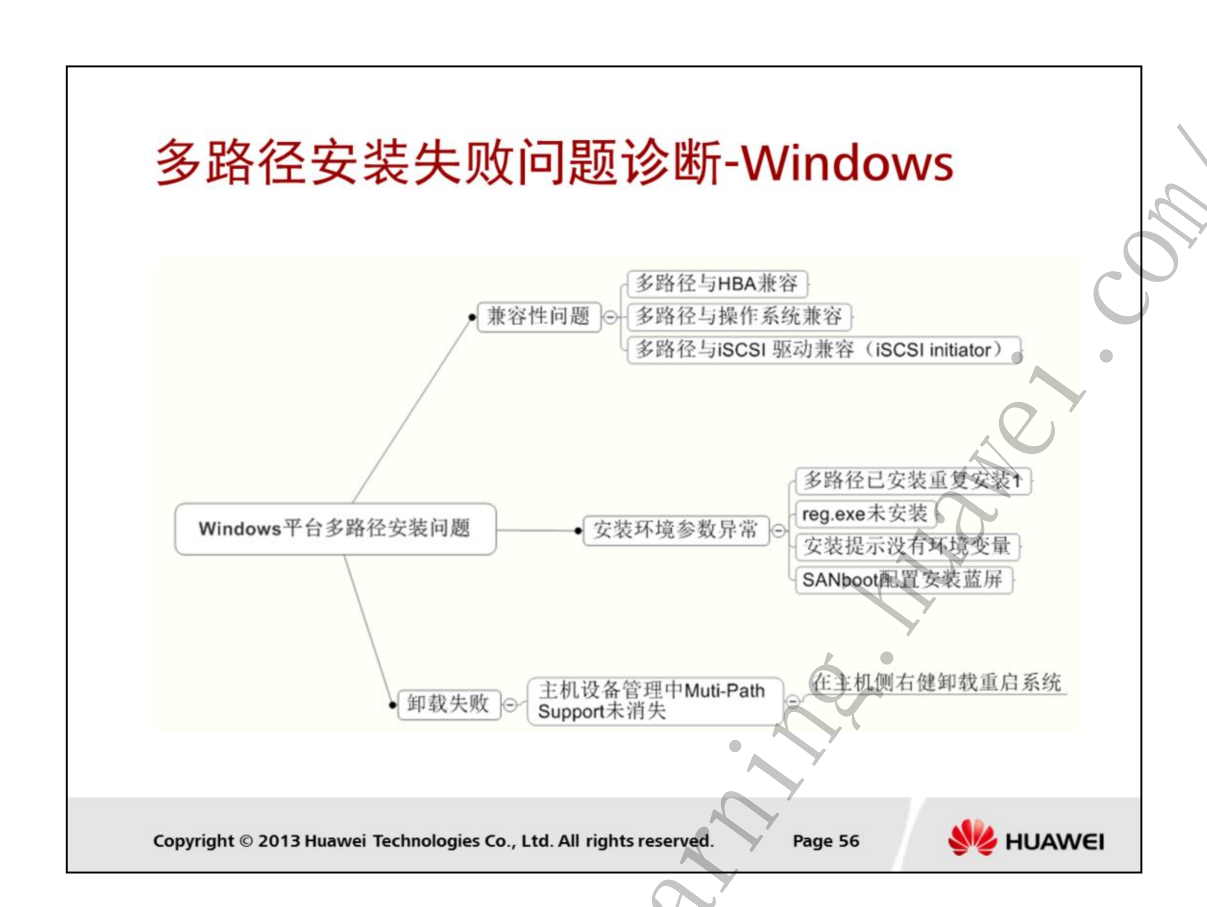


- UltraPath的其他初始化配置在安装前已经包含在安装包中，安装完成后即会采用默认配置进行工作。



目录




多路径安装失败问题诊断-Windows
Windows平台多路径安装问题
兼容性问题
多路径与HBA兼容
多路径与操作系统兼容
多路径与iSCSI 驱动兼容（iSCSI initiator）
安装环境参数异常
多路径已安装重复安装
reg.exe未安装
安装提示没有环境变量
SANboot配置安装蓝屏
卸载失败
主机设备管理中Muti-Path Support未消失
在主机侧右键卸载重启系统
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 56
HUAWEI

多路径安装失败问题处理-Windows

序号	问题	解决方案
1	多路径重复安装，安装时提示“已安装”	1. 在注册表中存在残留安装信息，搜索‘UltraPath’关键字的注册项，删除注册表中UltraPath相关的键值。 2. 删除当前安装程序中的UltraPath for Windows，重新安装
2	安装时提示reg.exe未安装	1. 系统文件reg.exe丢失，在安装盘SUPPORT\TOOLS下运行Setup.exe,资源工具包的安装，安装reg.exe注册表程序 2. 由于在资源工具包的安装过程中，程序自动将资源工具包的路径添加到WindowsXP的“PATH”变量下，因此安装完成后，用户可以直接在DOS命令行下运行reg.exe；
3	安装提示没有环境变量	可能出现用户手动修改环境变量时删除了系统的环境变量，在系统环境变量Path的值中添加“%SystemRoot%\system32;%SystemRoot%;”重新安装。

 Page 57 

- 说明：reg.exe的主要功能包括注册表项的查询、添加、删除、复制、保存、还原、加载、卸载、导入、导出以及注册表项目的比较和远程操作等。
- 建议：存储映射主机的LUN数量增加，安装多路径的时间也相对增长,建议在LUN数量较多的情况下，先安装多路径再进行主机映射。

多路径安装失败问题处理-Windows

序号	问题	解决方案
4	SANboot场景安装多路径出现蓝屏或其他环境出现蓝屏	1. 多数情况属于OS系统未安装SP1 /SP2补丁，先安装SP1/SP2补丁以后再安装多路径。 2. 如果系统已经安装SP1/SP2补丁，则需要分析蓝屏时产生的MEMERY.DMP文件，确认蓝屏发生的原因。



多路径安装失败问题诊断-Linux

多路径安装失败问题处理-Linux

序号	问题	解决方案
1	安装时提示多路径已安装	该主机已经安装了多路径，先卸载再安装。
2	使用光盘安装报错，提示找不到安装源（安装路径问题）	1. 将光盘中的多路径安装包拷贝到OS系统其它目录下。 2. 将小写的路径名称修改为正确路径名（区分大小写），再重新安装。
3	安装时提示安装rpm 包Error	检查多路径安装包格式是否为正确，执行dos2unix XXX.rpm 或者重新获取原始格式的多路径安装包。

 Page 60



- 该主机已经安装了多路径，先卸载再安装。
    1. 执行rpm -q UltraPath查询已安装的多路径版本。
    2. 卸载旧版本的多路径。例如，执行rpm -e UltraPath。
    3. 安装新版本的多路径。例如，执行rpm -ivhUltraPath*.rpm。
    4. 如果不能解决问题，执行rpm –e –allmatches XXX.rpm ,清除与多路径相关的文件，重新安装。

多路径安装失败问题处理-Linux

序号	问题	解决方案
4	多路径查询或卸载失败rpm管理器失败	可能出现rpm数据库损坏，执行rpm – rebuilddb，修复rpm数据库，重新执行查询或者卸载。
5	重启系统后多路径没有自动加载	1. 查看/boot/grub/menu.lst文件中，确定存在多路径启动选项“Linux with UltraPath”并且default的值指向该多路径启动选项“Linux with UltraPath”。如果不是，则先卸载掉多路径，修改/boot/grub/menu.lst，把default的指向值改为default 0，再重新安装多路径。 2. 确定/boot/grub/menu.lst文件中“Linux with UltraPath”启动项中“initrd (hd0,5)/boot/mpp-***-smp.img”中的文件“mpp-***-smp.img”在本地硬盘的/boot目录下存在。



多路径安装失败问题诊断-AIX

Page 62
HUAWEI

多路径安装失败问题处理-AIX

序号	问题	解决方案
1	安装失败，bin用户被屏蔽	1. 检查/etc/passwd，去掉“#bin:*:8:2::/bin:/usr/bin/ksh”的#。 2. 没有bin用户时，需手动添加bin用户。
2	bosboot验证未通过	跳过bosboot校验方式安装 1. 将/usr/sbin/下备份bosboot文件为bosboot.bak,修改bosboot在第二行添加 exit 0 2. 成功安装多路径以后，恢复bosboot文件。 参见案例。



多路径安装失败诊断总结

- 对于多路径安装失败，务必要有正确的诊断思路：
  - 首先确认是否存在兼容性问题。
  - 根据多路径安装失败提示，检查OS系统环境和配置参数。
  - 查看多路径安装日志信息，查看安装日志信息诊断。
  - 参考多路径升级指导书（一般在多路径软件包中有多路径升级指导），制定安装方案以及安装失败时的补救措施。



多路径安装失败案例

- 描述问题
  - AIX主机，安装多路径软件时提示如下错误:

    0503-409 installp: bosboot verification starting...

    0503-497 installp: An error occurred during bosboot verification processing.

    ERROR:install failed! please according to error info to check!
- 解决办法
  - 跳过bosboot校验方式安装，将/usr/sbin/下备份bosboot文件为bosboot.bak,修改bosboot，在第二行添加 exit 0，成功安装多路径以后，恢复bosboot文件。



- 原因分析
  - bosboot验证未通过，该命令用于保存磁盘的设备配置数据. “bosboot: /unix and /usr/lib/boot/unix must link to the same kernel file.” ,即/unix和/usr/lib/boot/unix指向的必须是同一内核文件，经检查” /unix “链接到“/usr/lib/boot/unix_64/unix” 。而 ” 文件/usr/lib/boot/unix_64/unix does not exist” .



目录



多路径运行过程问题诊断-Failover

- 多路径failover异常
  - 频繁failover操作
    - 检查同一个LUN映射给了多台主机
    - 检查多路径的链路物理状态
  - 链路中断，未执行failover操作
    - 另外一条链路异常
      - 检查HBA卡、交换端口、存储主机端口物理状态
      - 检查多路径服务和模块加载
      - 检查存储控制器状态
      - 检查故障路径控制器下LUN归属状态
    - /etc/iscsi.conf文件参数配置不合理
    - 检查open-iscsi驱动版本

多路径Failover失败处理步骤

序号	问题	解决方案
1	主机频繁Failover	同一LUN映射给了多台主机，导致多路径频切换，在集群和双机环境下允许LUN映射给主机组（映射给多台主机）。ISM上删除LUN对应的所有映射，将同一LUN只映射给一台主机。
2	iSCSI中断，未执行failover	1. /etc/iscsi.conf文件中配置不正确，检查/etc/iscsi.conf中是否有以下设置，默认情况没有。 ConnFailTimeout=1 2. open-iscsi为驱动版本不正确，更新open-iscsi为最新版本。



多路径Failover失败案例

- 描述问题
  - Linux主机，对LUN进行I/O操作时，映射给本主机的每个LUN上都有IO操作，A控制器上的网线断开，IO被挂住，归属A控的LUN未成功failover到B控。
- 分析原因
  - iSCSI的配置文件选项有错误，导致断线后iscsi启动器会不断重试。
- 解决办法
  - 在/etc/iscsi.conf文件中检查并添加配置信息:ConnFailTimeout=1



多路径运行过程问题诊断-Failback

Page 70

HUAWEI

多路径Failback异常分析和处理步骤

序号	问题	解决方案
1	主机频繁Failback	检查同一LUN映射给了多台主机，ISM上删除LUN对应的所有映射，将LUN只映射给一台主机。
2	控制器重新上电failback失败	上电的控制器没有向主机上报LUN信息，查看刚上电控制器上是否有路径，如果没有则执行hot_add扫盘操作。
3	LUN失效恢复failback失败	LUN失效恢复后，没有向主机上报LUN信息，在主机侧查看LUN的ReportedMissing的标志位，确认标志位为Y，执行hot_add扫盘操作。



多路径运行过程问题案例分享-Failback

- 描述问题
  - Linux主机SAS组网环境，A、B都有映射LUN，对所有LUN进行I/O操作。 复位A控制器，控制器下电后，IO failover到对端B，但A控制器重新上电后，IO一直没有failback。
- 解决办法
  - 确认A控制器上电成功后，在主机侧执行hot_add重新生成物理路径，发现存储设备LUN。

 Page 72



- 分析原因
  - SAS驱动使用了热插拔机制，当控制器下电后物理路径会被删除，但当控制器上电时由于物理链路的恢复和控制器完全上电有时间差，导致没有重新生成物理路径，多路径在规定的时间没有检测到正常恢复的物理路径，没有启动failback。

多路径运行过程问题诊断总结

- 对于多路径运行过程中问题，务必要有正确的诊断思路：
  - 首先确认主机侧物理LUN和虚拟LUN的个数是否正确。
  - 检查主机和存储设备之间LUN映射关系是否正确。
  - iSCSI或者HBA卡驱动版本是否正确。
  - 检查iSCSI的配置文件选项是否正确。
  - 收集日志信息诊断。



Window平台下Ultrapath状态查询命令

Windows 2003/2008下多路径查询命令	
命令	功能
upadm.exe verson	显示版本信息。
upadm.exe show diskMap	磁盘映射信息。
upadm.exe show array	服务器连接的阵列信息。
upadm.exe show arrayCtrl	服务器连接控制器信息。
upadm.exe show lunlo	虚拟LUNIO统计信息
upadm.exe show lunScsi	显示虚拟LUN详细信息。
upadm.exe show config	查询多路径配置信息

Windows 7/Solaris下多路径查询命令	
命令	功能
upadm help	显示帮助信息
upadm show version	查询版本信息
upadm show arrays	查询连接的阵列信息
upadm show vluns	所有映射的LUN信息。
upadm show vlun id=<ID1,ID2,...>	查询指定LUN的信息
upadm show luntrespass	查看当前LUN切换功能的状态



Linux平台下Ultrapath常用命令

Liunx常用命令列表	
命令	功能
up_esn	查看设备序列号信息
upadm help	查看简要帮助信息。
upadm chconfig	检查系统相关配置。
upadm show path	查看路径信息。
upadm show array	查看管理的阵列信息。
upadm show version	查看版本信息。
upadm show connectarray	查看连接到服务器上的所有存储系统信息
upadm show iostat array=array_id{lun=lun_id interval=seconds}	查看IO性能统计
upadm start hotscan	动态识别LUN
upadm start failback	手动启动Failback功能
upadm start forcerebalance	强制切换LUN的工作控制器为归属控制器



AIX平台下Ultrapath状态查询命令

AIX平台下Ultrapath状态查询命令	
命令	功能
upadm help	显示简要帮助信息。
upadm show version	查询软件版本信息。
upadm show daemon	查询守护进程运行状态。
upadm show option	查看可配置选项。
cfgmgr	扫描硬盘。
lspv	查看硬盘的概要信息。
upadm show lun [dev=updiskxx]	查看虚拟硬盘的详细信息。
lspath -F	查看硬盘的路径信息。
lspath	查看路径的优先级属性。





思考题

1. 主机在任何情况下都需要安装多路径软件吗?
2. UltraPath可以有哪几种负载均衡模式?





总结

- 主机操作系统与存储的连接
- UltraPath多路径在各操作系统的部署与管理





习题

- 判断题
  1. 主机HBA自带failover功能可以和Ultrapath的failover功能共同工作（T of F)
- 单选题
  1. AIX检查网卡IP地址使用命令是?（　　）

     A. lsdev -Cc if

     B. ifconfig –a

     C. smitty chinet

     D. smitty device



- 习题答案:
  - 判断题：1.F
  - 单选题：1.B

Thank you

www.huawei.com

HC120920005 统一存储系统故障诊断




HC120910005
统一存储系统故障诊断与排除
www.huawei.com
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



目标

- 学完本课程后，您将能够:
  - 熟悉统一存储系统故障处理原则，流程和方法
  - 熟悉SAN存储故障诊断思路和处理方法
  - 熟悉VIS存储故障诊断思路和处理方法





目录



故障诊断原则

先定外部，后定位内部	• 外部设备问题包括光纤、光缆、客户设备和断电等问题。 • 内部问题包括硬盘、控制器、接口模块等问题。
先高级别告警，后低级别告警	• 紧急：已经造成客户业务中断或者可能引起设备不可用的故障，需要立即采取措施修复。 • 重要：对设备产生部分影响或者对系统性能造成影响，需要采取纠正措施，以防止更严重的故障的发生。所有硬件类型的故障均属于该级别。 • 警告：对设备没有影响，系统检测到潜在的或者即将发生的影响业务的故障，且当前还没有影响业务。
先共性告警，后个别告警	• 在分析告警时，应先分析是个别问题还是共性问题，确定问题的影响范围。 • 需要确定是一个模块出问题，还是多个模块出现类似问题。

 Page 3 HUAWEI

- 先诊断外部因素、后诊断内部因素：诊断故障时，应先排除外部的可能因素，如电源中断、对接设备故障等。
- 先诊断网络、后诊断网元：根据网络拓扑图，分析网络环境是否正常、互连设备是否发生故障，尽可能准确定位出是网络中哪个环节发生故障。
- 先分析高级别告警、后分析低级别告警：在分析告警时，首先分析高级别的告警，如紧急告警、重要告警，然后分析低级别的告警，如提示告警。

故障诊断流程

- 故障诊断是指利用合理的方法，逐步找出产生故障的原因并解决故障，其基本思想是将可能的故障原因所构成的大集合缩减（或隔离）为若干个小的子集，从而使问题的复杂度下降。

开始
观察并记录指示灯状态
是否能登录ISM管理界面?
否
参考“无法登录ISM管理界面”进行故障排除
是
查看系统是否有事件信息?
否
观察故障现象并分析
定位故障产生的原因
根据定位出的原因进行故障排除
是
根据事件处理建议清除事件
故障是否排除?
是
否
收集故障信息
联系技术支持中心
清除事件后故障是否排除?
否
是
结束

HUAWEI

故障诊断是指利用合理的方法，逐步找出产生故障的原因并解决故障，其基本思想是将可能的故障原因所构成的大集合缩减（或隔离）为若干个小的子集，从而使问题的复杂度下降。

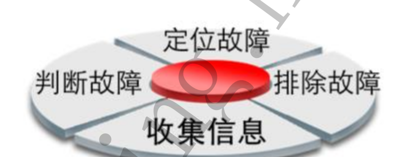
故障诊断前准备，步骤如下:

1. 登录ISM存储管理软件。
2. 确认SAN控制框部件的指示灯是否显示部件出现异常，指示灯的详细信息。
3. 如果刚在SAN控制框上更换了部件，并且SAN控制框无法运行，拆卸该部件并重新安装。

故障诊断流程包括收集信息、判断故障、定位故障和排除故障。

故障诊断步骤1---收集信息

- 全面、完整的故障信息有利于缩小判断故障的范围，加快判断、定位故障的速度和准确性，提高排除故障的效率。
- 故障诊断中信息收集主要考虑两方面的内容：
  - 收集对象：所需要收集的信息范围。
    - 硬件和软件版本
    - 故障时间、周期
    - 故障前的操作等
  - 收集途径：故障信息的来源。
    - 用户的故障申告
    - 巡检中发现的异常
    - 设备状态信息、日志信息

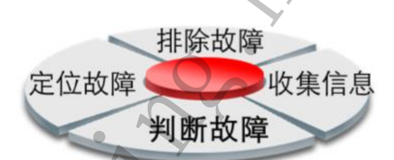


Page 5


- 收集的信息包括但不限于以下内容：
  - 故障现象、设备的硬件和软件版本
  - 故障时间、周期
  - 故障发生之前和故障发生时的操作
  - 故障发生时设备的输出信息、日志信息和告警信息
  - 故障发生后采取的措施
- 故障信息来源有下面几个：
  - 用户或客户服务中心的故障申告
  - 在日常维护或巡检中发现的异常（包括ISM查询到的告警信息）
  - 应用服务器的故障通告
  - 操作设备时获得的设备状态信息、日志信息
  - 通过命令行导出的调试信息

故障诊断步骤2---判断故障

- 在获取充分的故障信息后，初步判断故障的范围和种类。
  - 确定故障的范围
    - 所有业务发生故障
    - 部分业务发生故障
  - 确定故障的种类



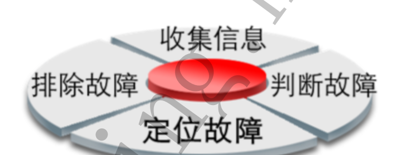
 Page 6



- 确定故障的范围
  - 确定控制框故障范围时需区分以下2种情况：
    - 所有业务发生故障：需进一步了解应用服务器是否同时发生故障。
    - 部分业务发生故障：需进一步了解故障业务类型及其在应用服务器上的分布情况，其他业务是否同时发生故障。
- 确定故障的种类
  - 确定故障的种类，即确定故障属于哪种类型。SAN控制框的故障种类通常根据其不同的功能模块进行划分。

故障诊断步骤3---定位故障

- 定位故障：
  - 指从众多可能原因中找出该单一原因的过程，通过分析、比较各种可能的故障原因，不断排除不可能因素，最终确定故障 发生的具体原因。
- 定位故障的重要性：
  - 准确而快速的定位有利于提高排除故障的效率，并有效避免因盲目操作设备而导致故障等级增大等人为事故的发生。



 Page 7



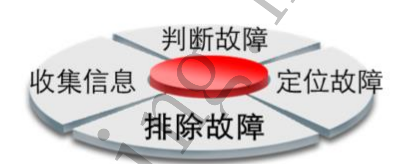
定位故障为排除故障提供指导和参考。

故障的产生在某一具体时刻具有单一性，这一特点决定了定位故障的基本方法。定位故障是指从众多可能原因中找出该单一原因的过程，通过分析、比较各种可能的故障原因，不断排除不可能因素，最终确定故障发生的具体原因。

准确而快速的定位有利于提高排除故障的效率，并有效避免因盲目操作设备而导致故障等级增大等人为事故的发生。

故障诊断步骤4---排除故障

- 排除故障：
  - 指采取适当的措施或步骤清除故障、恢复系统的过程。
- 基本方式：
  - 检修线路
  - 更换部件
  - 修改配置数据
  - 复位系统



 Page 8



排除故障是指采取适当的措施或步骤清除故障、恢复系统的过程，如检修线路、更换部件、修改配置数据、复位系统等。

排除故障过程中，请详细记录每一步操作以及产生的结果、现象。

故障定位的方法 – 告警信息分析法

- 当系统发生故障时，一般会伴随有大量的告警信息产生，通过查看告警信息并配合对性能数据的分析，可大概判断出所发生故障的类型和位置。
- 通过分析告警，可以定位故障的具体部位或原因，也可以配合其他方法定位故障原因。



应用场景

在能够正常收集到告警信息的情况下，告警分析法适用于定位任何故障。

故障定位的方法 – 替换法

- 替换法
  - 使用一个工作正常的部件去替换一个怀疑工作不正常的部件，从而达到定位故障、排除故障的目的。
  - 部件可以是一段光纤线、一根网线、一个控制器或者一个级联模块。
- 替换法的优势
  - 可以将故障定位到较细的位置
  - 对维护人员的要求不高，是一种比较实用的方法。

 Page 10 

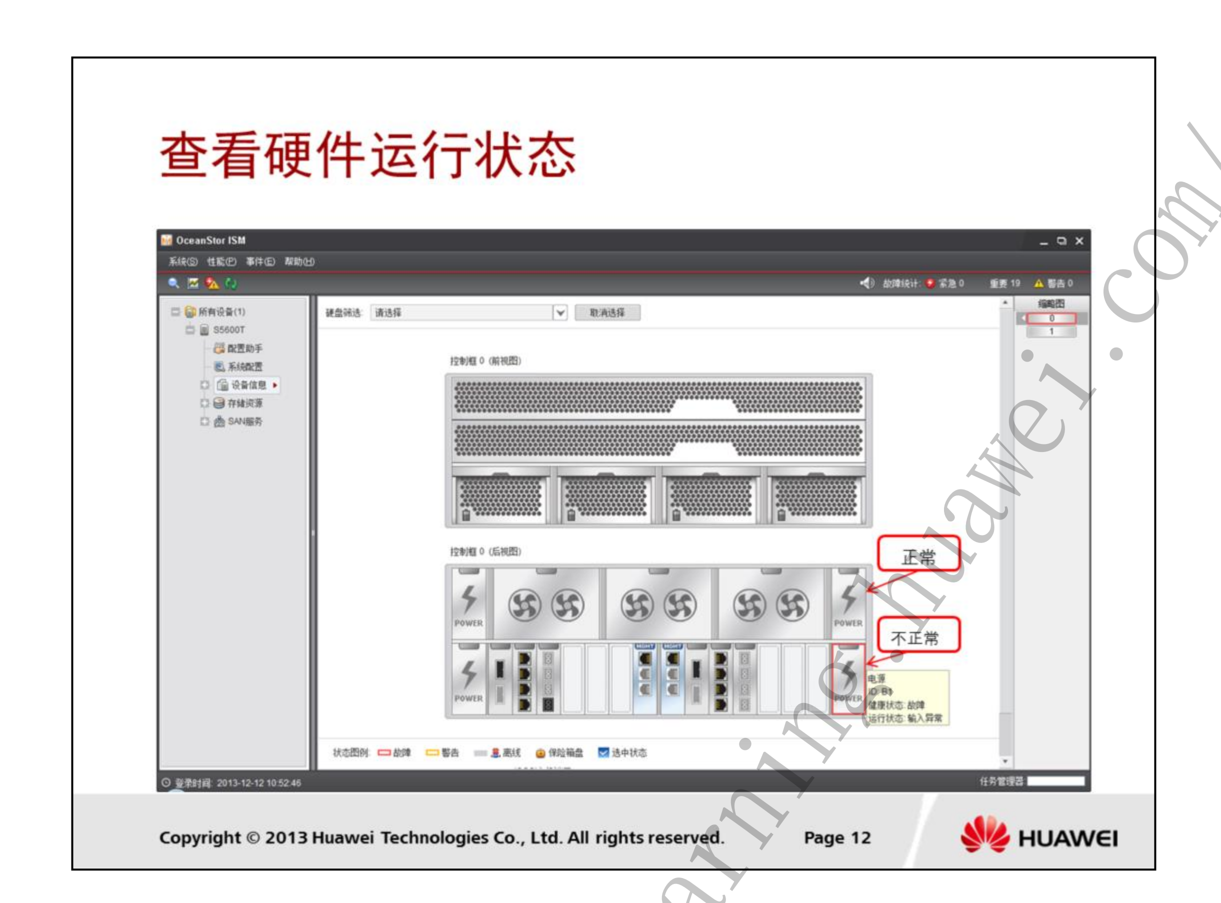
应用场景

替换法适用于硬件故障的处理，往往可以快速准确的定位出发生故障的部件，并且对维护人员没有特别的要求。使用替换法的局限在于事先必须准备相同的备件，因此要求进行较充分的前期准备工作。



目录





在ISM管理界面下的设备信息页面可以直观的查看各个硬件模块的运行状态，以上图为例：被红框圈定的电源模块的健康状态为故障。



在ISM管理界面的右上方可以看到各个告警级别的告警数量，直接点击相应的故障级别图标可以打开事件管理界面，在界面中可以看到所有详细告警信息。



在系统配置界面的导入导出部分可以对配置文件，运行数据和系统日志进行导出。

故障处理一般进程

- 观察并记录指示灯状态
- 登录ISM管理界面
- 根据事件处理建议处理事件
- 定位故障产生的原因
- 收集故障信息

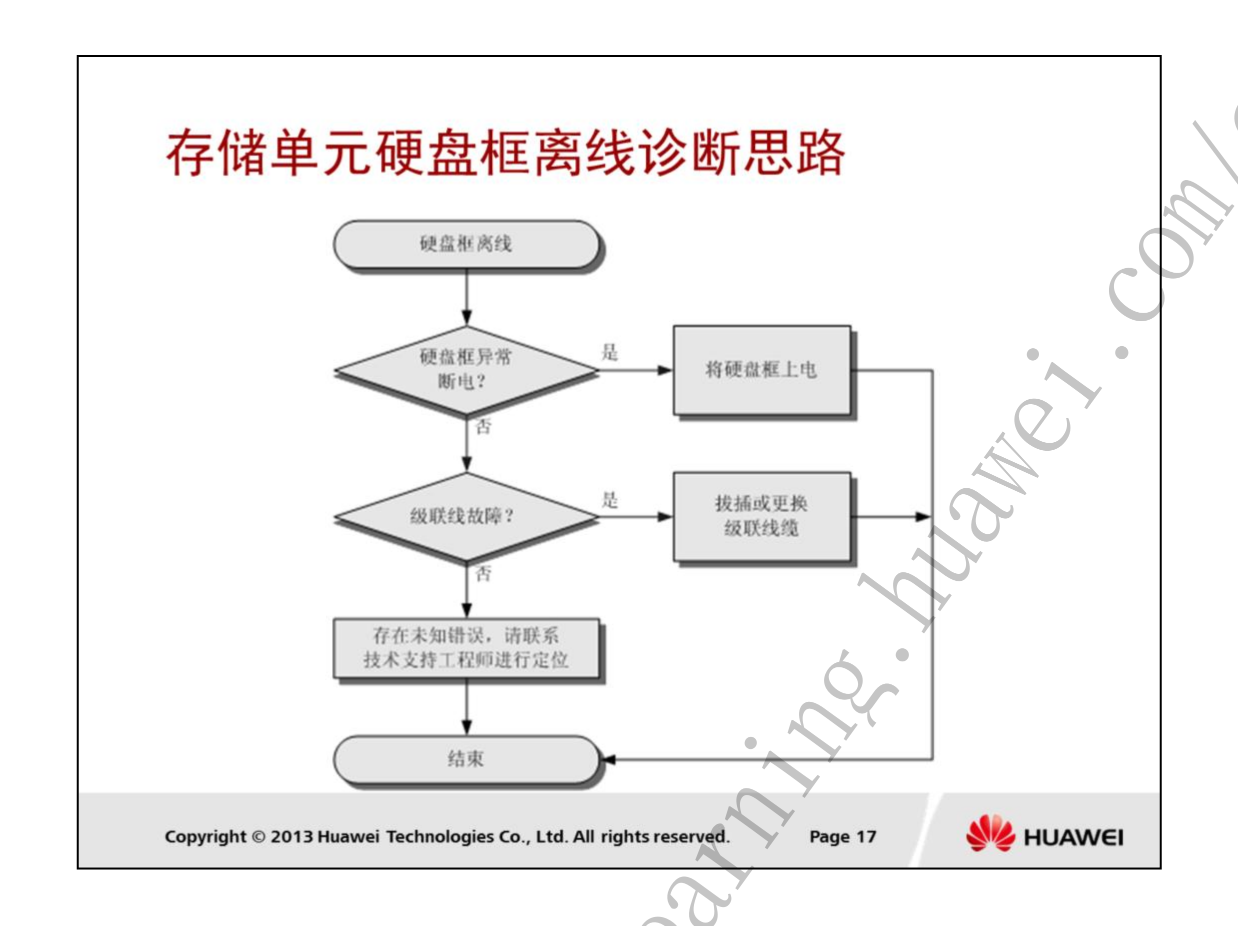
故障处理通常分为：了解故障概况、收集信息、分析原因、故障处理、验证恢复等几个进程。

- 观察并记录指示灯状态
  - 观察并记录指示灯的状态，通过指示灯的状态初步判断发生故障的模块。
  - 存储系统的各部件指示灯状态，请参见产品硬件描述
- 登录ISM管理界面
  - 登录ISM管理界面，可以查询存储系统的运行状态，以及是否有事件产生。
- 根据事件处理建议处理事件
  - 如果在ISM管理界面中查询到事件信息，则根据ISM管理界面提供的事件处理建议处理事件。
- 定位故障产生的原因
  - 定位故障是指从众多可能原因中找出故障发生的具体原因的过程。可通过分析、比较各种可能的故障原因，不断排除不可能因素，最终确定故障发生的具体原因。
- 收集故障信息
  - 遇到难以确定或解决的问题时，请您联系华为技术有限公司客户服务中心。同时，在向华为技术有限公司工程师反馈问题前，请注意收集故障处理所需的相关信息。

SAN存储阵列故障诊断

- 存储单元硬盘框离线
- 存储单元电源模块故障
- RAID组降级
- RAID组故障
- 虚拟快照故障
- LUN拷贝故障
- 资源LUN故障
- 应用服务器无法扫描到LUN
- FC链路异常
- iSCSI链路异常





硬盘框离线，将导致存储系统无法使用硬盘框上的存储资源，所有相关的功能和特性均不能使用。

- 现象描述

硬盘框已经通过级联线缆连接到控制框，登录ISM可以发现存储系统。但是在ISM导航树上，选择“所有设备 > SN_XX > 设备信息”（SN_XX表示正在管理的存储系统名称），在右侧的信息展示区选择存储单元，查看到级联的硬盘框有错误 01 。存储系统侧查看硬盘框的电源指示灯灭或mini SAS级联端口指示灯灭。

- 告警信息

在ISM菜单栏上，选择“事件 > 事件管理”，在弹出的“事件管理”对话框中，可以看到“硬盘框离线”的故障信息。

存储单元硬盘框离线处理步骤

序号	问题	解决方案
1	硬盘框异常断电	1.检查硬盘框电源指示灯是否亮绿色。 2.检查外部电源供电是否正常。 3.请与相应维护单位联系处理外部电源问题。处理完毕后，在ISM管理界面中是否可以看到级联的硬盘框。 4.拔插硬盘框电源线。 5.待硬盘框上电完成后，在ISM管理界面中是否可以看到级联的硬盘框。
2	级联线缆连接不正常或者级联线缆故障	1.观察存储系统上与硬盘框相连的级联端口link指示灯是否亮蓝色。 2.在ISM管理界面中，观察连接硬盘框的级联端口的“运行状态”是否显示为“未连接”。 3.拔插或更换级联线缆。 4.操作完成以后，相应级联端口link指示灯是否亮蓝色且在ISM管理界面中可以看到级联的硬盘框。

 Page 18



- 针对硬盘框异常断电

  1.检查硬盘框电源指示灯是否亮绿色。

  是 => 原因2。否 => 步骤2。

  2.检查外部电源供电是否正常。

  是 => 步骤4。否 => 步骤3。

  3.请与相应维护单位联系处理外部电源问题。处理完毕后，在ISM管理界面中是否可以看到级联的硬盘框。

  是 => 处理完毕。否 => 步骤4。

  4.拔插硬盘框电源线。

  5.待硬盘框上电完成后，在ISM管理界面中是否可以看到级联的硬盘框。

  是 => 处理完毕。否 => 原因2。

- 针对级联线缆连接不正常或者级联线缆故障

  1.观察存储系统上与硬盘框相连的级联端口link指示灯是否亮蓝色。

  是 => 保持故障环境并联系技术支持工程师进行处理。否 => 步骤2。

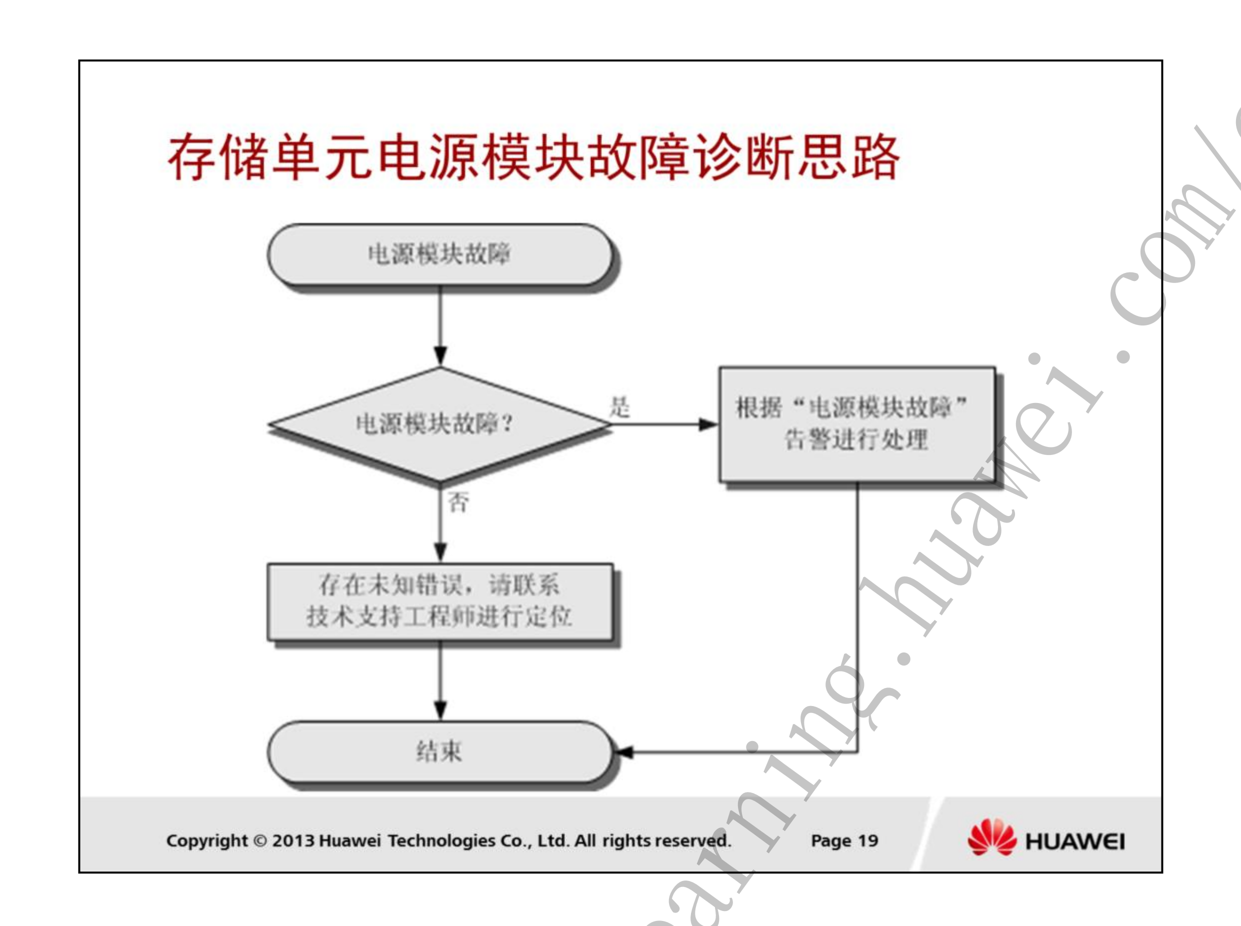
  2.在ISM管理界面中，观察连接硬盘框的级联端口的“运行状态”是否显示为“未连接”。

  是 => 步骤3。否 => 保持故障环境并联系技术支持工程师进行处理。

  3.拔插或更换级联线缆。

  4.操作完成以后，相应级联端口link指示灯是否亮蓝色且在ISM管理界面中可以看到级联的硬盘框。

  是 => 处理完毕。否 => 保持故障环境并联系技术支持工程师进行处理。



电源模块故障，可能会导致存储系统可靠性降低、读写性能下降等问题。

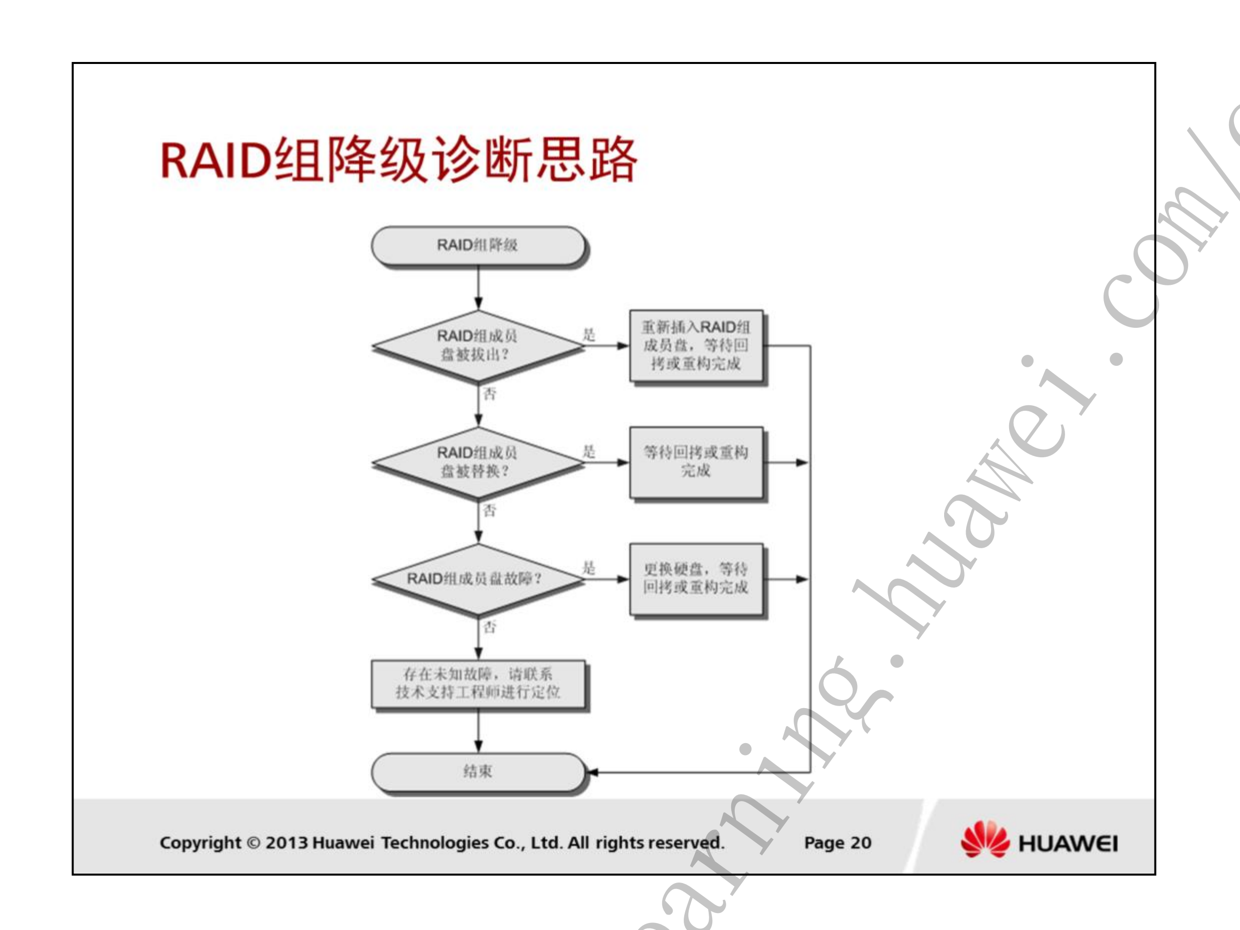
- 现象描述

在ISM导航树上，选择“所有设备 > SN_XX > 设备信息”（SN_XX表示存储系统的名称），在右侧的信息展示区选择存储单元，单击电源图标，在“电源属性”对话框中，可以看到电源的“健康状态”显示为“故障”。

在存储系统侧电源运行/告警指示灯亮红色。

- 告警信息

在ISM菜单栏上，选择“事件 > 事件管理”，在弹出的“事件管理”对话框中，可以看到电源模块故障信息。



RAID组降级可能会导致存储系统业务中断、读写性能下降、数据不一致或数据丢失等问题。

- 现象描述

在ISM导航树上，选择“所有设备 > SN_XX > 存储资源 > 块存储池”（SN_XX表示存储系统的名称），在右侧的信息展示区可以看到RAID组的“健康状态”显示为“降级”。

- 告警信息

在ISM菜单栏上，选择“事件 > 事件管理”，在弹出的“事件管理”对话框中，可以看到RAID组降级的故障信息。

RAID组降级处理步骤

序号	问题	解决方案
1	RAID组成员盘被拔出	1.在ISM管理界面上查看RAID组成员盘的“运行状态”是否显示为“离线”。 2.记录离线的RAID组成员盘位置。 3.根据步骤2中确认的成员盘位置，重新插入该硬盘并等待RAID组重构或回拷完成。 4.查看该成员盘的“运行状态”是否显示为“在线”且RAID组的“健康状态”显示为“正常”。
2	RAID组成员盘被其他硬盘替换	1.查看RAID组成员盘的“运行状态”是否显示为“正在回拷”或“正在重构”。 2.等待RAID组重构或回拷完成。 3.查看该成员盘的“运行状态”是否显示为“在线”且RAID组的“健康状态”显示为“正常”。

 Page 21



- 针对RAID组成员盘被拔出

1.在ISM管理界面上查看RAID组成员盘的“运行状态”是否显示为“离线”。

是 => 步骤2。否 => 原因2。

2.记录离线的RAID组成员盘位置。

3.根据步骤2中确认的成员盘位置，重新插入该硬盘并等待RAID组重构或回拷完成。

4.查看该成员盘的“运行状态”是否显示为“在线”且RAID组的“健康状态”显示为“正常”。

是 => 处理完毕。否 => 原因2。

- 针对RAID组成员盘被其他硬盘替换

1.查看RAID组成员盘的“运行状态”是否显示为“正在回拷”或“正在重构”。

是 => 步骤2。否 => 原因3。

2.等待RAID组重构或回拷完成。

3.查看该成员盘的“运行状态”是否显示为“在线”且RAID组的“健康状态”显示为“正常”。

是 => 处理完毕。否 => 原因3。

RAID组降级处理步骤

序号	问题	解决方案
3	RAID组成员盘故障	1.查看RAID组成员盘的“健康状态”是否显示为“未知”且“运行状态”显示为“故障”。 2.记录发生故障的RAID组成员盘位置。 3.更换该成员盘并等待RAID组重构或回拷完成。 4.查看该成员盘的“运行状态”是否显示为“在线”且RAID组的“健康状态”显示为“正常”。

Page 22



- 针对RAID组成员盘故障

1.查看RAID组成员盘的“健康状态”是否显示为“未知”且“运行状态”显示为“故障”。

是 => 步骤2。

否 => 保持故障环境并联系技术支持工程师进行处理。

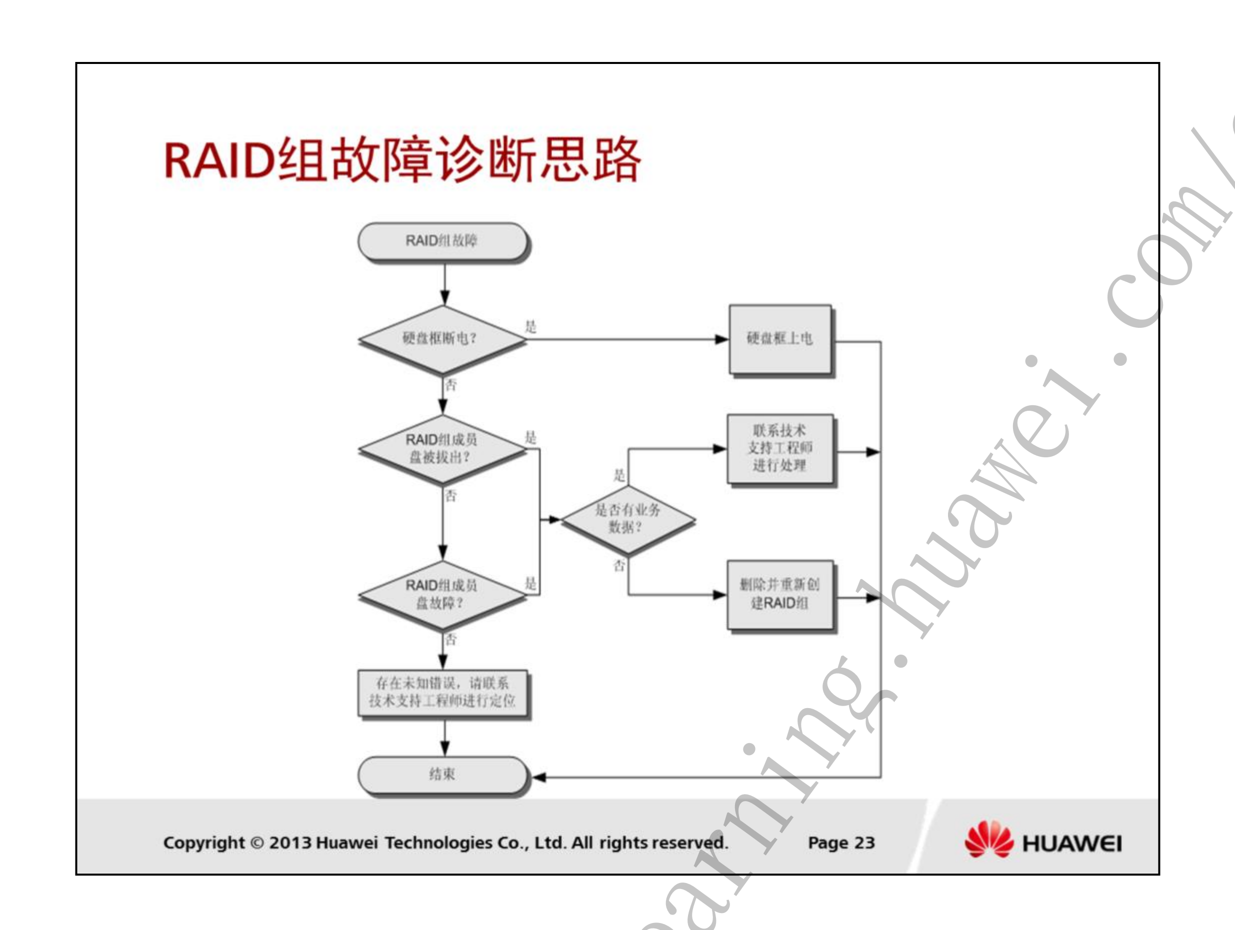
2.记录发生故障的RAID组成员盘位置。

3.更换该成员盘并等待RAID组重构或回拷完成。

4.查看该成员盘的“运行状态”是否显示为“在线”且RAID组的“健康状态”显示为“正常”。

是 => 处理完毕。

否 => 保持故障环境并联系技术支持工程师进行处理。



RAID组故障，将导致存储系统业务中断或数据丢失等问题。

- 现象描述

  在ISM导航树上，选择“所有设备 > SN_XX > 存储资源 > 块存储池”（SN_XX表示存储系统的名称），在右侧的信息展示区浏览RAID组信息，可以看到“健康状态”为“故障”且“运行状态”为“离线”的RAID组。

- 告警信息

  在ISM菜单栏上，选择“事件 > 事件管理”，在弹出的“事件管理”对话框中，可以看到RAID组的故障信息。

RAID组故障处理步骤

序号	问题	解决方案
1	硬盘框断电	1.在存储系统侧查看硬盘框前面板上的电源指示灯是否灭。 2.将硬盘框重新上电并等待RAID组状态自动恢复。 3.重新检查该RAID组的“运行状态”是否显示为“在线”。
2	RAID组成员盘有一个或多个被拔出	1.在ISM管理界面上查看RAID组成员盘的“运行状态”是否显示为“离线”。 2.RAID组中是否有业务数据。 3.删除故障的RAID组并重新创建RAID组。



- 针对硬盘框断电

  1.在存储系统侧查看硬盘框前面板上的电源指示灯是否灭。

  是 => 步骤2。否 => 原因2。

  2.将硬盘框重新上电并等待RAID组状态自动恢复。

  3.重新检查该RAID组的“运行状态”是否显示为“在线”。

  是 => 处理完毕。

  否 => 原因2。

- 针对RAID组成员盘有一个或多个被拔出

  1.在ISM管理界面上查看RAID组成员盘的“运行状态”是否显示为“离线”。

  是 => 步骤2。

  否 => 原因3。

  2.RAID组中是否有业务数据。

  是 => 保持故障环境并联系技术支持工程师进行处理。

  否 => 步骤3。

  3.删除故障的RAID组并重新创建RAID组。

RAID组故障处理步骤

序号	问题	解决方案
3	RAID组成员盘故障	1.查看RAID组中是否有两个或者两个以上（以RAID 5为例）成员盘的“健康状态”为“故障”。 2.RAID组中是否有业务数据。 3.删除故障的RAID组并重新创建RAID组。

 Page 25



- 针对RAID组成员盘故障

  1.查看RAID组中是否有两个或者两个以上（以RAID 5为例）成员盘的“健康状态”为“故障”。

  是 => 步骤2。

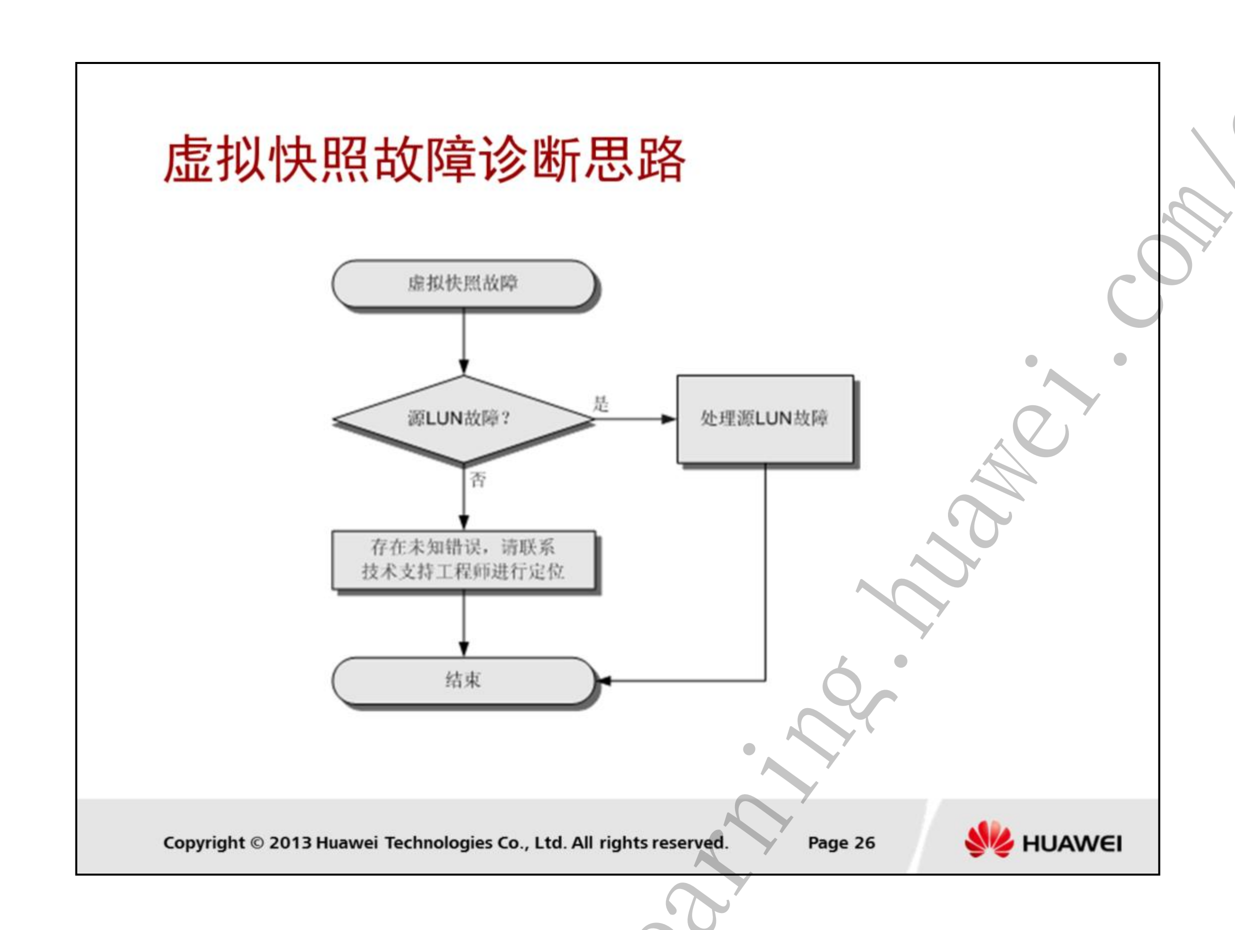
  否 => 保持故障环境并联系技术支持工程师进行处理。

  2.RAID组中是否有业务数据。

  是 => 保持故障环境并联系技术支持工程师进行处理。

  否 => 步骤3。

  3.删除故障的RAID组并重新创建RAID组。



虚拟快照故障，可能会导致快照数据丢失或快照不可用。

- 现象描述

在ISM导航树上，选择“所有设备 > SN_XX > SAN服务 > 快照”（SN_XX表示存储系统的名称），在右侧的信息展示区浏览虚拟快照的“健康状态”，可以看到处于“故障”状态的虚拟快照。

- 告警信息

在ISM菜单栏上，选择“事件 > 事件管理”，在弹出的“事件管理”对话框中，可以看到快照相关故障信息。

虚拟快照故障处理步骤

序号	问题	解决方案
1	虚拟快照的源LUN故障	1.在ISM上查看虚拟快照的源LUN的“健康状态”是否显示为“故障”。 2.在ISM上查看RAID组的“健康状态”是否显示为“故障”。 3.根据RAID组故障进行处理。 4.操作完成后，查看虚拟快照的“健康状态”是否显示为“正常”。

 Page 27



- 针对虚拟快照的源LUN故障

  1.在ISM上查看虚拟快照的源LUN的“健康状态”是否显示为“故障”。

    是 => 步骤2。

    否 => 保持故障环境并联系技术支持工程师进行处理。

  2.在ISM上查看RAID组的“健康状态”是否显示为“故障”。

    是 => 步骤3。

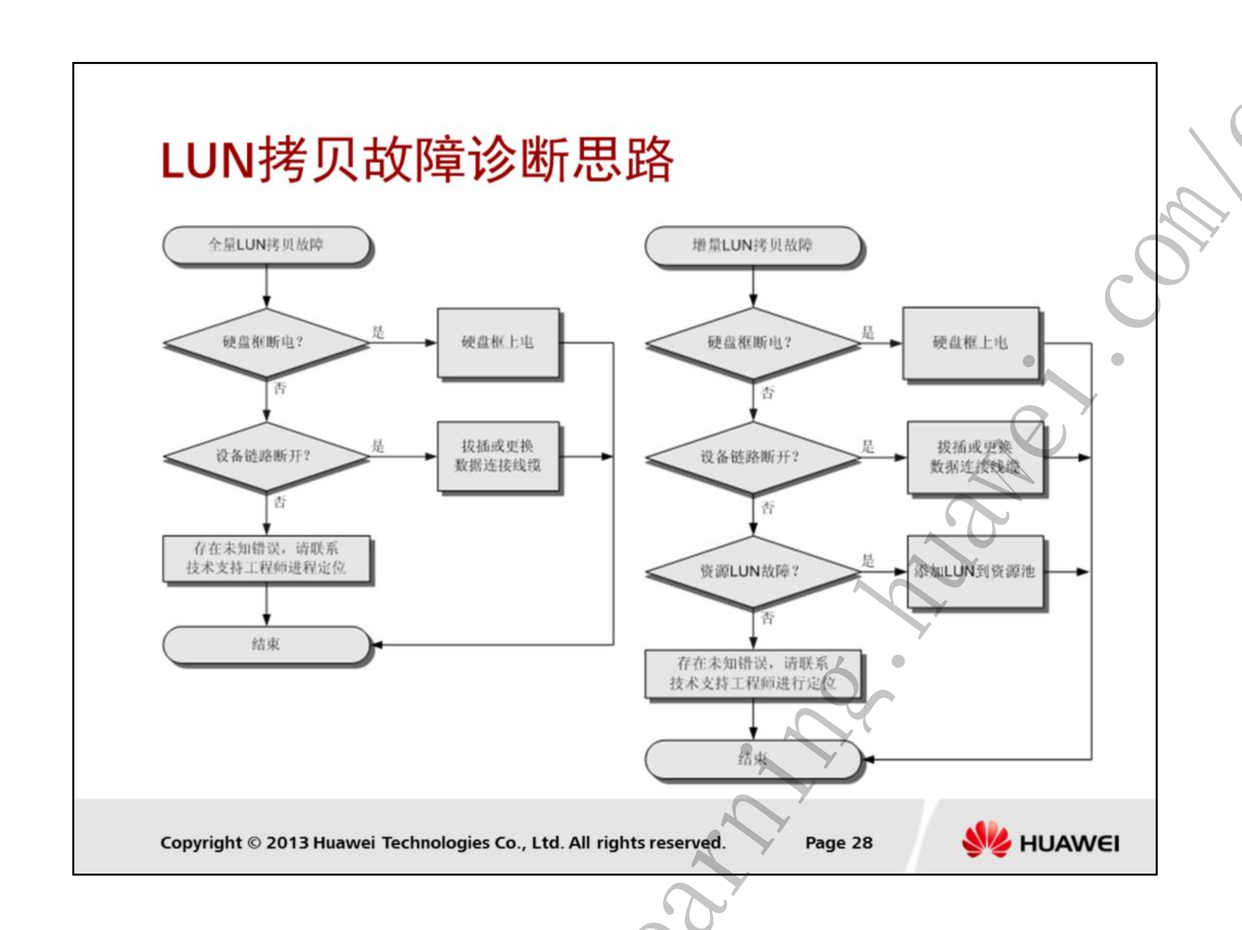
    否 => 保持故障环境并联系技术支持工程师进行处理。

  3.根据RAID组故障进行处理。

  4.操作完成后，查看虚拟快照的“健康状态”是否显示为“正常”。

    是 => 处理完毕。

    否 => 保持故障环境并联系技术支持工程师进行处理。



LUN拷贝故障，将导致所有与LUN拷贝相关的操作无法正常执行。

- 现象描述

在ISM导航树上，选择“所有设备 > SN_XX > SAN服务 > LUN拷贝”（SN_XX表示存储系统的名称），在右侧的信息展示区浏览LUN拷贝的“健康状态”，可以看到处于“故障”状态的LUN拷贝。

- 告警信息

在ISM菜单栏上，选择“事件 > 事件管理”，在弹出的“事件管理”对话框中，可以看到LUN拷贝相关故障信息。

LUN拷贝故障处理步骤

序号	问题	解决方案
1	源LUN或目标LUN所在的硬盘框异常断电	1.在存储系统侧检查源LUN或目标LUN所在硬盘框电源运行指示灯状态是否灭。 2.检查外部电源供电是否正常。 3.请与相应维护单位联系处理外部电源问题。处理完毕后，查看LUN拷贝的“健康状态”是否显示为“正常”。 4.拔插硬盘框电源线。待硬盘框上电完成后，查看LUN拷贝的“健康状态”是否显示为“正常”。



- 针对源LUN或目标LUN所在的硬盘框异常断电

  1.在存储系统侧检查源LUN或目标LUN所在硬盘框电源运行指示灯状态是否灭。

    是 => 步骤2。

    否 => 原因2。

  2.检查外部电源供电是否正常。

    是 => 步骤4。

    否 => 步骤3。

  3.请与相应维护单位联系处理外部电源问题。处理完毕后，查看LUN拷贝的“健康状态”是否显示为“正常”。

    是 => 处理完毕。

    否 => 步骤4。

  4.拔插硬盘框电源线。待硬盘框上电完成后，查看LUN拷贝的“健康状态”是否显示为“正常”。

    是 => 处理完毕。

    否 => 原因2。

LUN拷贝故障处理步骤

序号	问题	解决方案
2	设备间链路断开（适用于设备间的LUN拷贝）	1.在存储系统侧观察设备上相应级联端口或主机端口指示灯状态是否灭。 2.确认该故障的LUN拷贝是增量LUN拷贝还是全量LUN拷贝。 3.查看连接硬盘框的级联端口或设备间的链路两端的主机端口的“运行状态”是否显示为“未连接”。 4.拔插或更换断开的数据连接线缆。操作完成以后，相应级联端口或主机端口指示灯是否亮绿色，且可以在ISM管理界面中看到级联端口或主机端口的“运行状态”显示为“连接”。 5.查看故障的LUN拷贝的“健康状态”是否显示为“正常”。 6.该故障的LUN拷贝是增量LUN拷贝还是全量LUN拷贝。

Page 30


- 针对设备间链路断开（适用于设备间的LUN拷贝）

  1.在存储系统侧观察设备上相应级联端口或主机端口指示灯状态是否灭。

  是 => 步骤3。否 => 步骤2。

  2.确认该故障的LUN拷贝是增量LUN拷贝还是全量LUN拷贝。

  全量LUN拷贝 => 保持故障环境并联系技术支持工程师进行处理。

  增量LUN拷贝 => 原因3。

  3.查看连接硬盘框的级联端口或设备间的链路两端的主机端口的“运行状态”是否显示为“未连接”。

  是 => 步骤4。否 => 步骤5。

  4.拔插或更换断开的数据连接线缆。

  操作完成以后，相应级联端口或主机端口指示灯是否亮绿色，且可以在ISM管理界面中看到级联端口或主机端口的“运行状态”显示为“连接”。

  是 => 步骤5。否 => 保持故障环境并联系技术支持工程师进行处理。

  5.查看故障的LUN拷贝的“健康状态”是否显示为“正常”。

  是 => 处理完毕。否 => 步骤6。

  6.该故障的LUN拷贝是增量LUN拷贝还是全量LUN拷贝。

  全量LUN拷贝 => 保持故障环境并联系技术支持工程师进行处理。

  增量LUN拷贝 => 原因3。

LUN拷贝故障处理步骤

序号	问题	解决方案
3	源LUN所在设备上的资源LUN故障（适用于增量LUN拷贝）	1.在ISM上，查看源LUN所在设备上的是否有资源LUN的“健康状态”显示为“故障”。 2.根据资源LUN故障进行处理。 3.操作完成后，请重新创建一个或多个LUN并添加到资源池。添加的LUN的总容量应该大于等于故障的资源LUN容量。 4.操作完成后，查看故障的LUN拷贝的“健康状态”是否显示为“正常”。

Page 31


- 针对源LUN所在设备上的资源LUN故障（适用于增量LUN拷贝）

  1.在ISM上，查看源LUN所在设备上的是否有资源LUN的“健康状态”显示为“故障”。

    是 => 步骤2。

    否 => 保持故障环境并联系技术支持工程师进行处理。

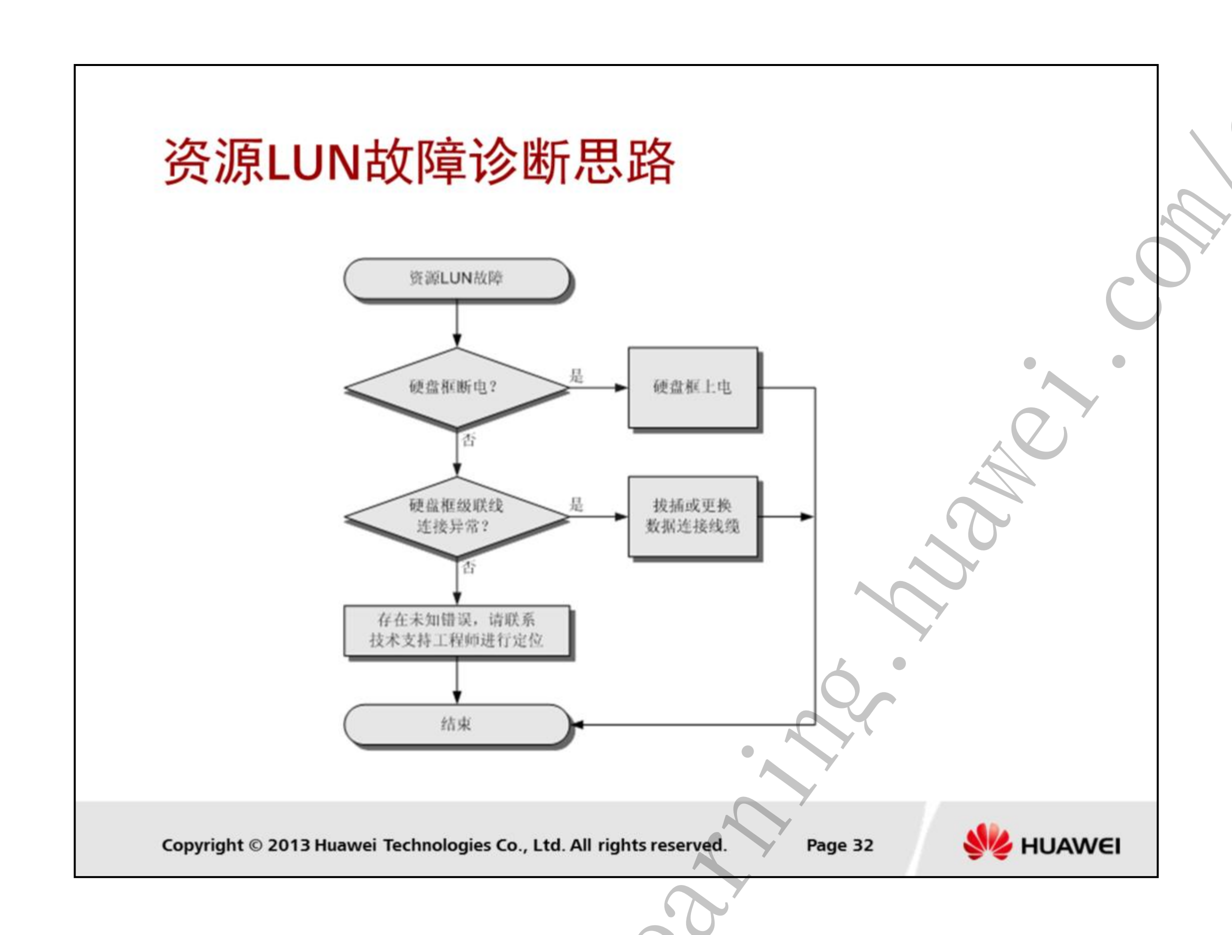
  2.根据资源LUN故障进行处理。

  3.操作完成后，请重新创建一个或多个LUN并添加到资源池。添加的LUN的总容量应该大于等于故障的资源LUN容量。

  4.操作完成后，查看故障的LUN拷贝的“健康状态”是否显示为“正常”。

    是 => 处理完毕。

    否 => 保持故障环境并联系技术支持工程师进行处理。



资源LUN故障，可能导致增量LUN拷贝、虚拟快照、异步远程复制等增值功能异常，从而导致数据丢失，影响系统性能和可靠性。

- 现象描述

在ISM导航树上，选择“所有设备 > SN_XX > 存储资源 > 资源池”（SN_XX表示存储系统的名称），在右侧的信息展示区中选择资源池A或B，浏览信息展示区下方的资源LUN“健康状态”，可以看到处于“故障”状态的资源LUN。

- 告警信息

在ISM菜单栏上，选择“事件 > 事件管理”，在弹出的“事件管理”对话框中，可以看到资源LUN相关的故障信息。

资源LUN故障处理步骤

序号	问题	解决方案
1	资源LUN所在的硬盘框异常断电	1.在存储系统侧检查资源LUN所在的硬盘框电源运行指示灯是否灭。 2.检查外部电源供电是否正常。 3.请与相应维护单位联系处理外部电源问题。处理完毕后，查看资源LUN的“健康状态”是否显示为“正常”。 4.拔插硬盘框电源线。 待硬盘框上电完成后，查看资源LUN的“健康状态”是否显示为“正常”。



- 针对资源LUN所在的硬盘框异常断电

  1.在存储系统侧检查资源LUN所在的硬盘框电源运行指示灯是否灭。

  是 => 步骤2。

  否 => 保持故障环境并联系技术支持工程师进行处理。

  2.检查外部电源供电是否正常。

  是 => 步骤4。

  否 => 步骤3。

  3.请与相应维护单位联系处理外部电源问题。处理完毕后，查看资源LUN的“健康状态”是否显示为“正常”。

  是 => 处理完毕。

  否 => 步骤4。

  4.拔插硬盘框电源线。

  待硬盘框上电完成后，查看资源LUN的“健康状态”是否显示为“正常”。

  是 => 处理完毕。

  否 => 原因2。

资源LUN故障处理步骤

序号	问题	解决方案
2	资源LUN所在的硬盘框级联线连接异常	1.在存储系统侧观察设备上相应级联端口link指示灯状态是否灭。 2.在ISM上，查看资源LUN所在的硬盘框的“运行状态”是否显示为“离线”。 3.拔插或更换断开的数据传输线缆。 4.操作完成后，查看相应级联端口link指示灯是否亮蓝色，且在ISM管理界面中可以看到级联端口“运行状态”显示为“连接”。 5.查看资源LUN的“健康状态”是否显示为“正常”。

 Page 34



- 针对资源LUN所在的硬盘框级联线连接异常

  1.在存储系统侧观察设备上相应级联端口link指示灯状态是否灭。

    是 => 步骤2。

    否 => 保持故障环境并联系技术支持工程师进行处理。

  2.在ISM上，查看资源LUN所在的硬盘框的“运行状态”是否显示为“离线”。

    是 => 步骤3。

    否 => 保持故障环境并联系技术支持工程师进行处理。

  3.拔插或更换断开的数据传输线缆。

  4.操作完成后，查看相应级联端口link指示灯是否亮蓝色，且在ISM管理界面中可以看到级联端口“运行状态”显示为“连接”。

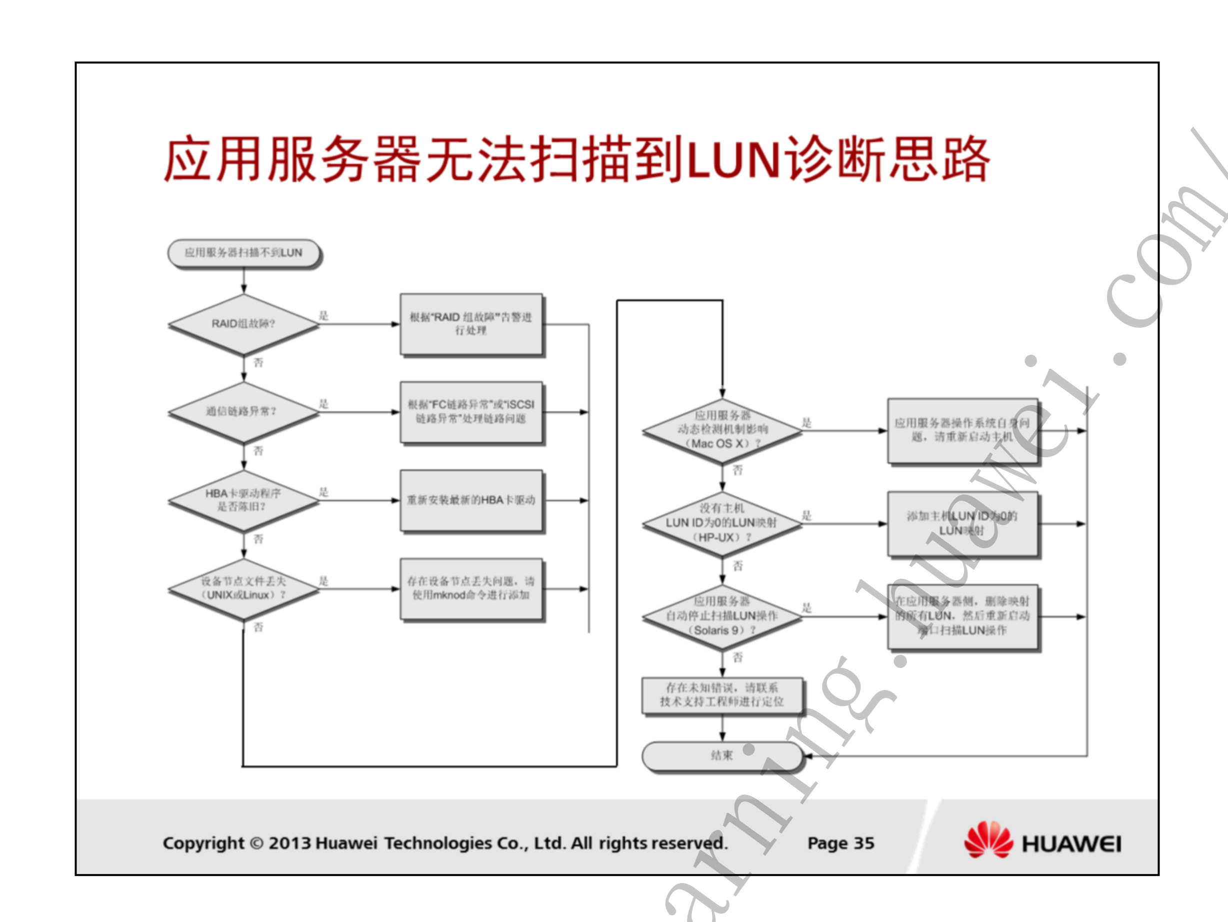
    是 => 步骤5。

    否 => 保持故障环境并联系技术支持工程师进行处理。

  5.查看资源LUN的“健康状态”是否显示为“正常”。

    是 => 处理完毕。

    否 => 保持故障环境并联系技术支持工程师进行处理。



应用服务器无法扫描到LUN，将导致应用服务器无法使用存储系统提供的资源。

- 现象描述

  当用户为应用服务器映射了LUN后，在应用服务器侧无法发现映射的LUN。

应用服务器无法扫描到LUN处理步骤

序号	问题	解决方案
1	RAID组故障	1.请检查存储系统中是否存在“0xE01F90002 RAID组故障”的告警。 2.根据“0xE01F90002 RAID组故障”告警处理RAID组故障。 3.在应用服务器上重新扫描LUN，查看故障是否解决。
2	链路通信异常	1.在ISM管理界面上查看与应用服务器连接的FC或iSCSI主机端口的“运行状态”是否显示为“未连接”。 2.确认是iSCSI组网还是FC组网。 3.根据FC链路连接异常处理FC组网的链路问题，处理完成后转至步骤5。 4.根据iSCSI链路连接异常处理iSCSI组网的链路问题，处理完成后转至步骤5。 5.在应用服务器上重新扫描LUN，查看故障是否解决。



- 针对RAID组故障

1.请检查存储系统中是否存在“0xE01F90002 RAID组故障”的告警。

是 => 步骤2。否 => 原因2。

2.根据“0xE01F90002 RAID组故障”告警处理RAID组故障。

3.在应用服务器上重新扫描LUN，查看故障是否解决。

是 => 处理完毕。否 => 原因2。

- 针对链路通信异常

1.在ISM管理界面上查看与应用服务器连接的FC或iSCSI主机端口的“运行状态”是否显示为“未连接”。

是 => 步骤2。否 => 原因3。

2.确认是iSCSI组网还是FC组网。

FC组网 => 步骤3。iSCSI组网 => 步骤4。

3.根据FC链路连接异常处理FC组网的链路问题，处理完成后转至步骤5。

4.根据iSCSI链路连接异常处理iSCSI组网的链路问题，处理完成后转至步骤5。

5.在应用服务器上重新扫描LUN，查看故障是否解决。

是 => 处理完毕。否 => 原因3。

应用服务器无法扫描到LUN处理步骤

序号	问题	解决方案
3	HBA卡的驱动程序版本过于陈旧	1.请检查应用服务器上安装的HBA卡驱动程序是否为陈旧的版本。 2.在应用服务器上卸载已安装的HBA卡驱动程序，然后重新安装最新版本的HBA卡驱动程序。 3.在应用服务器上重新扫描LUN，查看故障是否解决。
4	设备节点文件丢失（适用于UNIX或Linux）	1.确认应用服务器的操作系统是否为Linux或UNIX。 2.在应用服务器侧查看“dev”目录下，是否有对应的设备节点文件，例如“dev/sdb”。 3.在应用服务器的“Terminal”中执行命令mknod，创建设备节点。 命令mknod一般使用的形式为：mknod Name {b | c} Major Minor，其中Name代表设备名称，b | c代表块设备或者字符设备，Major代表主设备号，Minor代表次设备号。 4.在应用服务器上重新扫描LUN，查看故障是否解决。

 Page 37



- 针对HBA卡的驱动程序版本过于陈旧

  1.请检查应用服务器上安装的HBA卡驱动程序是否为陈旧的版本。

  是 => 步骤2。否 => 原因4。

  2.在应用服务器上卸载已安装的HBA卡驱动程序，然后重新安装最新版本的HBA卡驱动程序。

  3.在应用服务器上重新扫描LUN，查看故障是否解决。

  是 => 处理完毕。否 => 原因4。

- 针对设备节点文件丢失（适用于UNIX或Linux）

  1.确认应用服务器的操作系统是否为Linux或UNIX。

  是 => 步骤2。否 => 保持故障环境并联系技术支持工程师进行处理。

  2.在应用服务器侧查看“dev”目录下，是否有对应的设备节点文件，例如“dev/sdb”。

  是 => 原因5。否 => 步骤3。

  3.在应用服务器的“Terminal”中执行命令mknod，创建设备节点。

  命令mknod一般使用的形式为：mknod Name {b ｜ c} Major Minor，其中Name代表设备名称，b ｜ c代表块设备或者字符设备，Major代表主设备号，Minor代表次设备号。

  4.在应用服务器上重新扫描LUN，查看故障是否解决。

  是 => 处理完毕。否 => 原因5。

应用服务器无法扫描到LUN处理步骤

序号	问题	解决方案
5	应用服务器动态检测机制影响（适用于Mac OS X）	1.确认应用服务器的操作系统是否为Mac OS X。 2.请重新启动Mac OS X应用服务器。 当Mac OS X应用服务器没有LUN映射时，Mac OS X应用服务器本身的动态检测机制可能无法生效，所以需要重新启动应用服务器来触发该动态机制。 3.在应用服务器上重新扫描LUN，查看故障是否解决。
6	应用服务器没有“主机LUN ID”为0的LUN映射（适用于HP-UX）	1.确认应用服务器的操作系统是否为HP-UX。 2.在ISM上查看HP-UX应用服务器的LUN映射，是否存在“主机LUN ID”为0的LUN映射。 3.为HP-UX应用服务器添加“主机LUN ID”为0的LUN映射。 4.在应用服务器上重新扫描LUN，查看故障是否解决。

 Page 38



- 应用服务器动态检测机制影响（适用于Mac OS X）

  1.确认应用服务器的操作系统是否为Mac OS X。

  是 => 步骤2。否 => 原因6。

  2.请重新启动Mac OS X应用服务器。

  当Mac OS X应用服务器没有LUN映射时，Mac OS X应用服务器本身的动态检测机制可能无法生效，所以需要重新启动应用服务器来触发该动态机制。

  3.在应用服务器上重新扫描LUN，查看故障是否解决。

  是 => 处理完毕。否 => 保持故障环境并联系技术支持工程师进行处理。

- 应用服务器没有“主机LUN ID”为0的LUN映射（适用于HP-UX）

  1.确认应用服务器的操作系统是否为HP-UX。

  是 => 步骤2。否 => 原因7。

  2.在ISM上查看HP-UX应用服务器的LUN映射，是否存在“主机LUN ID”为0的LUN映射。

  是 => 保持故障环境并联系技术支持工程师进行处理。否 => 步骤3。

  3.为HP-UX应用服务器添加“主机LUN ID”为0的LUN映射。

  4.在应用服务器上重新扫描LUN，查看故障是否解决。

  是 => 处理完毕。否 => 保持故障环境并联系技术支持工程师进行处理。

应用服务器无法扫描到LUN处理步骤

序号	问题	解决方案
7	应用服务器自动停止扫描LUN操作（适用于Solaris 9）	1.确认应用服务器的操作系统是否为Solaris 9。 2.确认Solaris 9应用服务器是否安装了“SAN Foundation Software”。 3.运行cfgadm al命令查询端口的WWN（World Wide Name）号。 4.运行cfgadm al -o show_FCP_dev c2::WWN命令重新启动端口扫描LUN操作。其中show_FCP_dev表示显示LUN信息，c2表示步骤3中查询到的端口号，WWN表示步骤3中查询到的WWN号。 查看故障是否解决。

 Page 39



- 针对应用服务器自动停止扫描LUN操作（适用于Solaris 9）

1.确认应用服务器的操作系统是否为Solaris 9。

是 => 步骤2。

否 => 保持故障环境并联系技术支持工程师进行处理。

2.确认Solaris 9应用服务器是否安装了“SAN Foundation Software”。

是 => 步骤3。

否 => 保持故障环境并联系技术支持工程师进行处理。

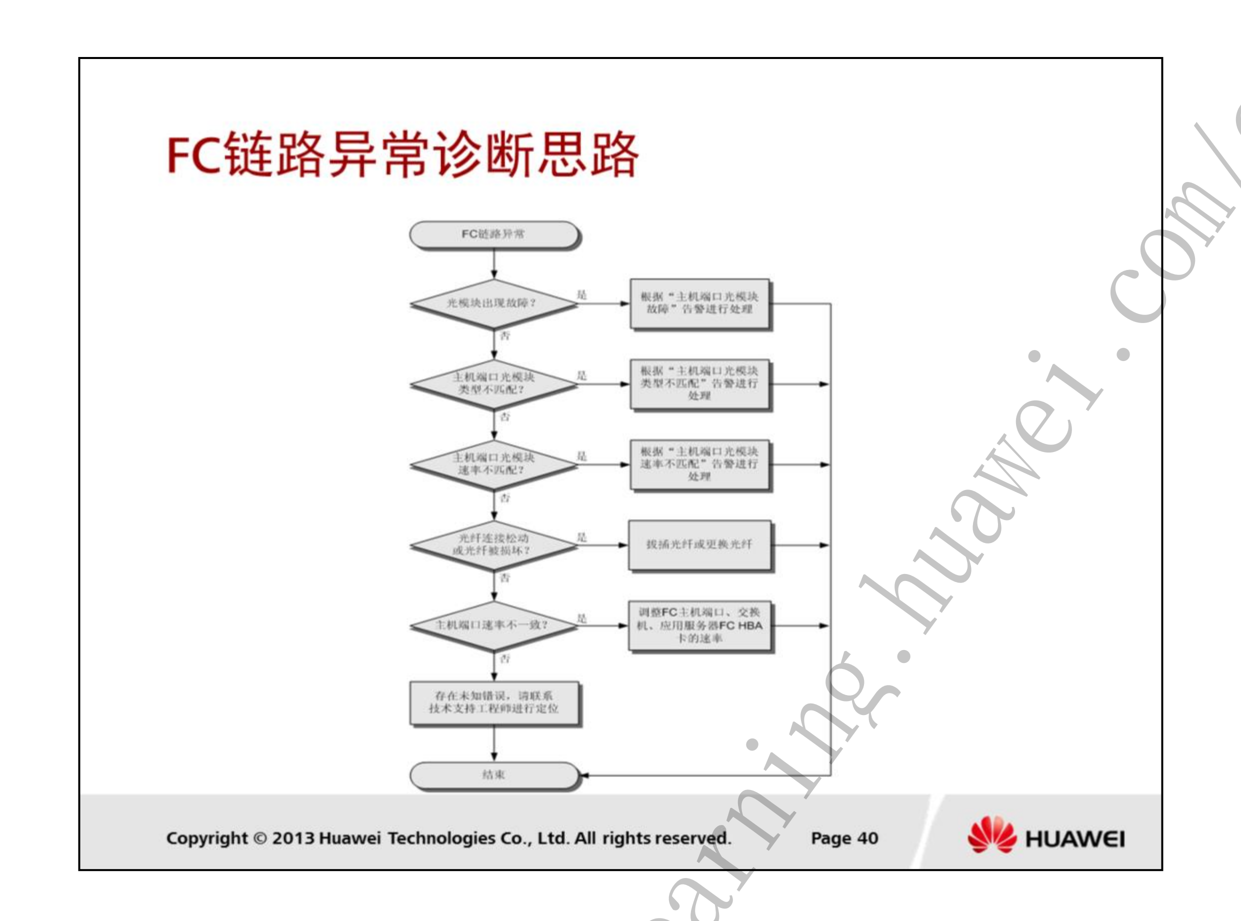
3.运行cfgadm al命令查询端口的WWN（World Wide Name）号。

4.运行cfgadm al -o show_FCP_dev c2::WWN命令重新启动端口扫描LUN操作。其中show_FCP_dev表示显示LUN信息，c2表示步骤3中查询到的端口号，WWN表示步骤3中查询到的WWN号。

查看故障是否解决。

是 => 处理完毕。

否 => 保持故障环境并联系技术支持工程师进行处理。



FC链路异常，可能会导致应用服务器与存储系统之间产生业务中断、数据丢失等问题。

- 现象描述

FC接口模块上的端口已通过光纤与应用服务器建立了物理连接。在ISM导航树上，选择“所有设备 > SN_XX > 设备信息 > 存储单元 > 端口”（SN_XX表示存储系统的名称），在右侧的信息展示区浏览FC主机端口信息，可以看到“健康状态”为“--”且“运行状态”为“未连接”的FC端口。在设备现场发现该FC主机端口link/speed指示灯亮红灯或灭。

- 告警信息

登录ISM，在ISM导航树上选择“事件 > 事件管理”，在弹出的“事件管理”对话框的“故障列表”页签下，可能存在主机端口和光模块相关告警。

本文以“0xE020B0003 主机端口光模块故障”告警（适用于S2200T/S2600T/S5500T/S5600T/S5800T/S6800T）为例，如果存在“0xE020B000B 主机端口光模块故障”告警（适用于S5500T），则以“0xE020B000B 主机端口光模块故障”中的处理步骤进行处理。

本文以“0xE020B0004 主机端口光模块类型不匹配”告警（适用于S2200T/S2600T/S5500T/S5600T/S5800T/S6800T）为例，如果存在“0xE020B000C 主机端口光模块类型不匹配”告警（适用于S5500T），则以“0xE020B000C 主机端口光模块类型不匹配”中的处理步骤进行处理。

本文以“0xE020B0005 主机端口光模块速率不匹配”告警（适用于S2200T/S2600T/S5500T/S5600T/S5800T/S6800T）为例，如果存在“0xE020B000D 主机端口光模块速率不匹配”告警（适用于S5500T），则以“0xE020B000D 主机端口光模块速率不匹配”中的处理步骤进行处理。

FC链路异常处理步骤

序号	问题	解决方案
1	光模块出现故障	1.在ISM上查看告警信息，是否存在“0xE020B0003 主机端口光模块故障”告警（适用于T系列）或“0xE020B000B 主机端口光模块故障”告警（适用于S5500T）。 2.根据“0xE020B0003 主机端口光模块故障”告警（适用于T系列）或“0xE020B000B 主机端口光模块故障”告警（适用于S5500T）进行处理。 3.操作结束后，FC主机端口的link指示灯是否亮绿色或蓝色，且在ISM中该主机端口的“运行状态”显示为“已连接”。
2	主机端口光模块类型不匹配	1.在ISM上查看告警信息，是否存在“0xE020B0004 主机端口光模块类型不匹配”（适用于T系列）告警或“0xE020B000C 主机端口光模块类型不匹配”告警（适用于S5500T）。 2.根据“0xE020B0004 主机端口光模块类型不匹配”告警（适用于T系列）或“0xE020B000C 主机端口光模块类型不匹配”告警（适用于S5500T）进行处理。 3.操作结束后，FC主机端口的link指示灯是否亮绿色或蓝色，且在ISM中该主机端口的“运行状态”显示为“已连接”。

 Page 41



- 针对光模块出现故障

  1.在ISM上查看告警信息，是否存在“0xE020B0003 主机端口光模块故障”告警（适用于T系列）或“0xE020B000B 主机端口光模块故障”告警（适用于S5500T）。

  是 => 步骤2。否 => 原因2。

  2.根据“0xE020B0003 主机端口光模块故障”告警（适用于T系列）或“0xE020B000B 主机端口光模块故障”告警（适用于S5500T）进行处理。

  3.操作结束后，FC主机端口的link指示灯是否亮绿色或蓝色，且在ISM中该主机端口的“运行状态”显示为“已连接”。

  是 => 处理完毕。否 => 原因4。

- 针对主机端口光模块类型不匹配

  1.在ISM上查看告警信息，是否存在“0xE020B0004 主机端口光模块类型不匹配”（适用于T系列）告警或“0xE020B000C 主机端口光模块类型不匹配”告警（适用于S5500T）。

  是 => 步骤2。否 => 原因3。

  2.根据“0xE020B0004 主机端口光模块类型不匹配”告警（适用于T系列）或“0xE020B000C 主机端口光模块类型不匹配”告警（适用于S5500T）进行处理。

  3.操作结束后，FC主机端口的link指示灯是否亮绿色或蓝色，且在ISM中该主机端口的“运行状态”显示为“已连接”。

  是 => 处理完毕。否 => 原因4。

FC链路异常处理步骤

序号	问题	解决方案
3	主机端口光模块速率不匹配	1.在ISM上查看告警信息，是否存在“0xE020B0005 主机端口光模块速率不匹配”告警（适用于S2200T/S2600T/S5500T/S5600T/S5800T/S6800T）或“0xE020B000D 主机端口光模块速率不匹配”告警（适用于S5500T）。 2.根据“0xE020B0005 主机端口光模块速率不匹配”告警（适用于S2200T/S2600T/S5500T/S5600T/S5800T/S6800T）或“0xE020B000D 主机端口光模块速率不匹配”告警（适用于S5500T）进行处理。 3.操作结束后，FC主机端口的link指示灯是否亮绿色或蓝色，且在ISM中该主机端口的“运行状态”显示为“已连接”。
4	光纤连接松动或光纤被损坏	1.拔插光纤或更换光纤。 2.操作完成后，FC主机端口的link指示灯是否亮绿色或蓝色，且在ISM中该主机端口的“运行状态”显示为“已连接”。

 Page 42



- 针对主机端口光模块速率不匹配

  1.在ISM上查看告警信息，是否存在“0xE020B0005 主机端口光模块速率不匹配”告警（适用于S2200T/S2600T/S5500T/S5600T/S5800T/S6800T）或“0xE020B000D 主机端口光模块速率不匹配”告警（适用于S5500T）。

  是 => 步骤2。否 => 原因4。

  2.根据“0xE020B0005 主机端口光模块速率不匹配”告警（适用于S2200T/S2600T/S5500T/S5600T/S5800T/S6800T）或“0xE020B000D 主机端口光模块速率不匹配”告警（适用于S5500T）进行处理。

  3.操作结束后，FC主机端口的link指示灯是否亮绿色或蓝色，且在ISM中该主机端口的“运行状态”显示为“已连接”。

  是 => 处理完毕。否 => 原因4。

- 针对光纤连接松动或光纤被损坏

  1.拔插光纤或更换光纤。

  2.操作完成后，FC主机端口的link指示灯是否亮绿色或蓝色，且在ISM中该主机端口的“运行状态”显示为“已连接”。

  是 => 处理完毕。否 => 原因5。

FC链路异常处理步骤

序号	问题	解决方案
5	主机端口速率不一致	1.现场组网环境是直连组网还是交换机组网。 交换机组网 => 步骤2。直连组网 => 步骤8。 2.查看FC主机端口速率与存储系统连接的交换机端口速率是否一致。 关于如何检查交换机端口或FC HBA卡速率请咨询对应的厂商或查看其说明书。 3.调整FC主机端口的“配置速率”，使FC主机端口的“配置速率”与对应的交换机端口速率一致。 4.调整速率后，FC主机端口的link指示灯是否亮绿色或蓝色，且在ISM中该主机端口的“运行状态”显示为“已连接”。 5.查看应用服务器HBA卡工作速率与对应交换机端口的速率是否一致。 6.调整与应用服务器连接的交换机端口速率，使交换机端口速率与应用服务器FC HBA卡速率一致。

 Page 43 

- 针对主机端口速率不一致

1.现场组网环境是直连组网还是交换机组网。

交换机组网 => 步骤2。直连组网 => 步骤8。

2.查看FC主机端口速率与存储系统连接的交换机端口速率是否一致。

关于如何检查交换机端口或FC HBA卡速率请咨询对应的厂商或查看其说明书。

是 => 步骤5。否 => 步骤3。

3.调整FC主机端口的“配置速率”，使FC主机端口的“配置速率”与对应的交换机端口速率一致。

4.调整速率后，FC主机端口的link指示灯是否亮绿色或蓝色，且在ISM中该主机端口的“运行状态”显示为“已连接”。

是 => 处理完毕。否 => 步骤5。

5.查看应用服务器HBA卡工作速率与对应交换机端口的速率是否一致。

是 => 保持故障环境并联系技术支持工程师进行处理。

否 => 步骤6。

6.调整与应用服务器连接的交换机端口速率，使交换机端口速率与应用服务器FC HBA卡速率一致。

FC链路异常处理步骤

序号	问题	解决方案
5	主机端口速率不一致	7.调整速率后，FC主机端口的link指示灯是否亮绿色或蓝色，且在ISM中该主机端口的“运行状态”显示为“已连接”。 8.查看FC主机端口速率与存储系统连接的应用服务器的HBA卡的工作速率是否一致。 9.调整FC主机端口的“配置速率”，使FC主机端口的“配置速率”与对应的应用服务器FC HBA卡设置的工作速率一致。 10.调整速率后，FC主机端口的link指示灯是否亮绿色或蓝色，且在ISM中该主机端口的“运行状态”显示为“已连接”。



- 针对主机端口速率不一致

7.调整速率后，FC主机端口的link指示灯是否亮绿色或蓝色，且在ISM中该主机端口的“运行状态”显示为“已连接”。

是 => 处理完毕。

否 => 保持故障环境并联系技术支持工程师进行处理。

8.查看FC主机端口速率与存储系统连接的应用服务器的HBA卡的工作速率是否一致。

是 => 保持故障环境并联系技术支持工程师进行处理。

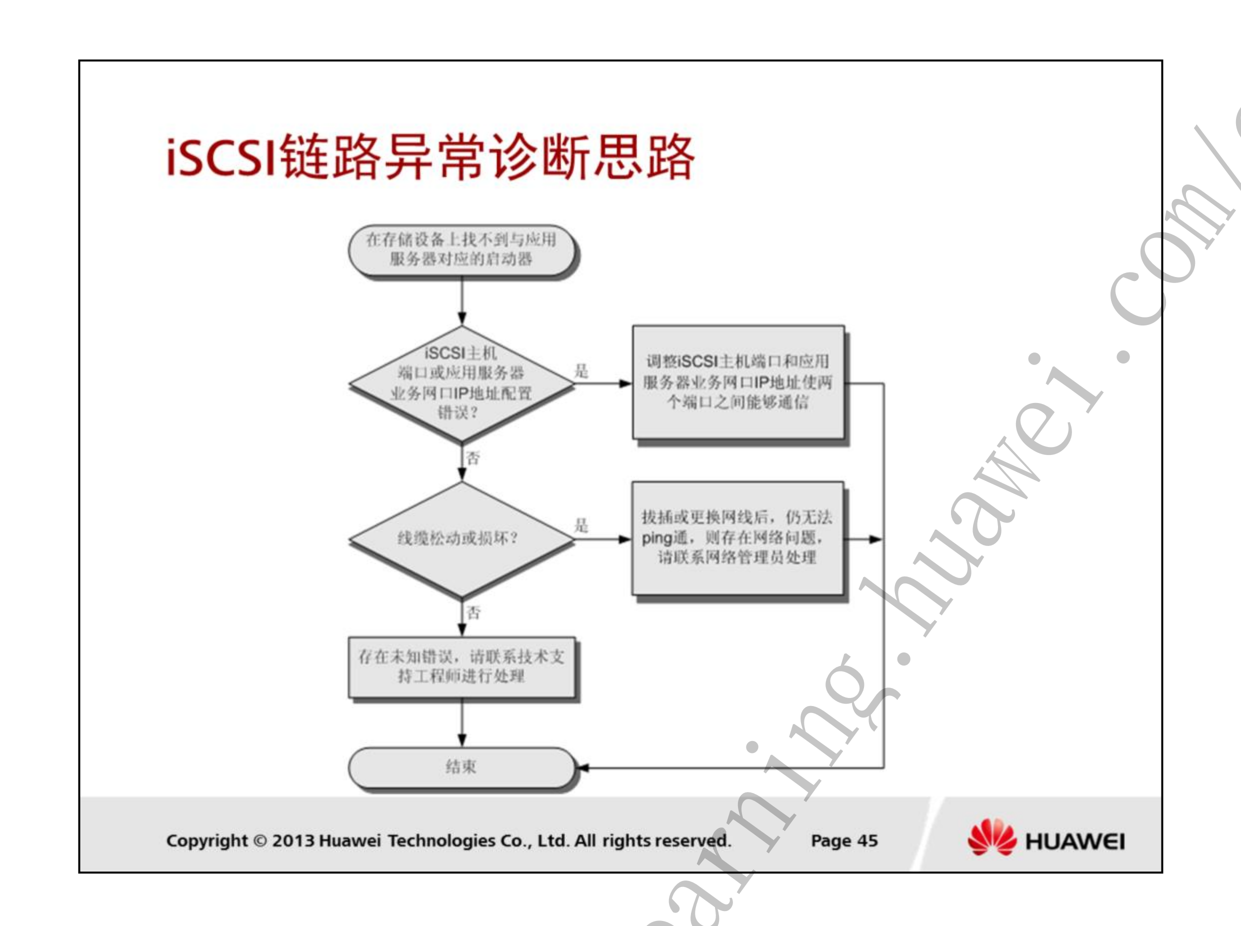
否 => 步骤9。

9.调整FC主机端口的“配置速率”，使FC主机端口的“配置速率”与对应的应用服务器FC HBA卡设置的工作速率一致。

10.调整速率后，FC主机端口的link指示灯是否亮绿色或蓝色，且在ISM中该主机端口的“运行状态”显示为“已连接”。

是 => 处理完毕。

否 => 保持故障环境并联系技术支持工程师进行处理。



iSCSI链路异常，可能会导致应用服务器与存储系统之间产生业务中断、数据丢失等问题。

- 现象描述

在ISM导航树上，选择“所有设备 > SN_XX > 设备信息 > 存储单元 > 端口”（SN_XX表示存储系统的名称），在右侧的信息展示区浏览iSCSI主机端口信息，可以看到“健康状态”为“--”并且“运行状态”为“未连接”的iSCSI端口。在设备现场发现该iSCSI主机端口link指示灯灭。

- 告警信息

登录ISM，在ISM导航树上选择“事件 > 事件管理”，在弹出的“事件管理”对话框的“故障列表”页签下，可能存在“主机端口连接断开”的告警。

iSCSI链路异常处理步骤

序号	问题	解决方案
1	iSCSI主机端口IP地址或应用服务器业务网口IP地址配置错误	1.确认在应用服务器上是否可以ping通iSCSI主机端口IP地址。 2.确认现场组网环境是直连组网还是交换机组网。 直连组网 => 步骤3。交换机组网 => 步骤4。 3.修改iSCSI主机端口IP地址，使iSCSI主机端口IP地址与应用服务器业务网口IP地址在同一个网段，然后转至步骤5。 您也可以在应用服务器上修改应用服务器的业务网口IP地址，使其与iSCSI主机端口IP地址在同一个网段上。 4.分别为iSCSI主机端口和应用服务器添加路由，使iSCSI主机端口和应用服务器能够通信，然后转至步骤5。 5.请在应用服务器上运行ping命令查看网络链路是否可以通，其中目的地址为存储系统iSCSI主机端口IP地址。

 Page 46



- 针对iSCSI主机端口IP地址或应用服务器业务网口IP地址配置错误

1.确认在应用服务器上是否可以ping通iSCSI主机端口IP地址。

是 => 保持故障环境并联系技术支持工程师进行处理。

否 => 步骤2。

2.确认现场组网环境是直连组网还是交换机组网。

直连组网 => 步骤3。交换机组网 => 步骤4。

3.修改iSCSI主机端口IP地址，使iSCSI主机端口IP地址与应用服务器业务网口IP地址在同一个网段，然后转至步骤5。

您也可以在应用服务器上修改应用服务器的业务网口IP地址，使其与iSCSI主机端口IP地址在同一个网段上。

4.分别为iSCSI主机端口和应用服务器添加路由，使iSCSI主机端口和应用服务器能够通信，然后转至步骤5。

5.请在应用服务器上运行ping命令查看网络链路是否可以通，其中目的地址为存储系统iSCSI主机端口IP地址。

是 => 处理完毕。否 => 原因2。

iSCSI链路异常处理步骤

序号	问题	解决方案
2	应用服务器与存储系统之间的线缆松动或损坏	1.拔插或更换连接存储系统与应用服务器之间的网线。 2.操作结束后，请在应用服务器上运行ping命令查看网络链路是否可以通，其中目的地址为存储系统iSCSI主机端口IP地址。 3.操作结束后，iSCSI主机端口的link指示灯是否亮绿色或蓝色，且在ISM中该主机端口的“运行状态”显示为“连接”。

 Page 47



- 针对应用服务器与存储系统之间的线缆松动或损坏

  1.拔插或更换连接存储系统与应用服务器之间的网线。

  2.操作结束后，请在应用服务器上运行ping命令查看网络链路是否可以通，其中目的地址为存储系统iSCSI主机端口IP地址。

  是 => 步骤3。

  否 => 保持故障环境并联系技术支持工程师进行处理。

  3.操作结束后，iSCSI主机端口的link指示灯是否亮绿色或蓝色，且在ISM中该主机端口的“运行状态”显示为“连接”。

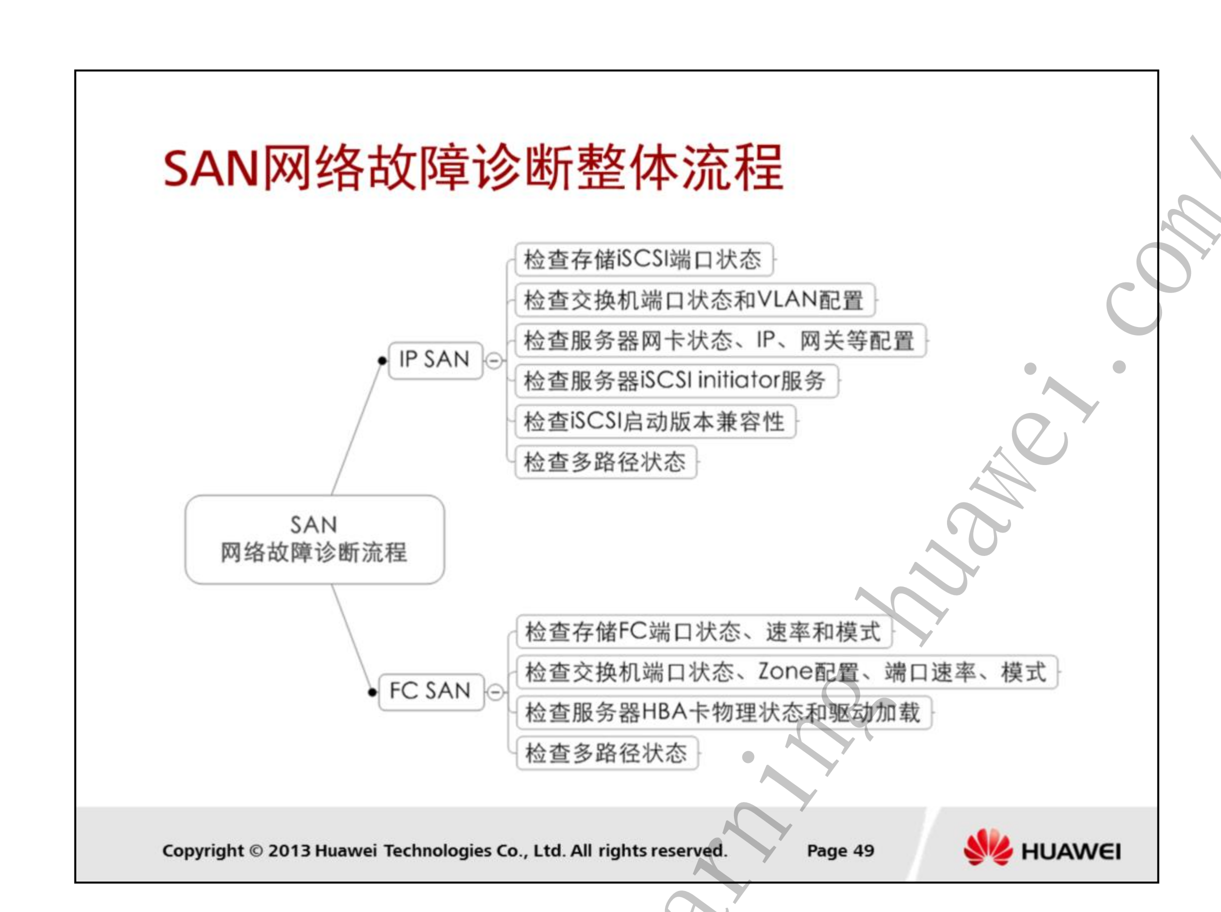
  是 => 处理完毕。

  否 => 保持故障环境并联系技术支持工程师进行处理。

SAN网络故障分类和定义

- 按故障对象不同分：
  - IP SAN：1）IP无法连通；2）端口限制；3）VLAN设置不合理；4）交换机端口故障；5）网卡故障；
  - FC SAN：1）端口模式不匹配；2）端口速率不匹配；3）兼容性；4）SFP光功率不足；5）光纤或端口物理故障；6）HBA卡驱动加载异常；
- 按故障性质分：
  - 物理故障：硬件设备或者物理设备出现异常，无法正常工作；
  - 逻辑故障：设备逻辑状态异常；例如驱动或模块加载异常；
  - 配置异常：链路通信参数（速率、模式）配置不合理导致异常；




SAN网络故障诊断整体流程
SAN
网络故障诊断流程
IP SAN
检查存储iSCSI端口状态
检查交换机端口状态和VLAN配置
检查服务器网卡状态、IP、网关等配置
检查服务器iSCSI initiator服务
检查iSCSI启动版本兼容性
检查多路径状态
FC SAN
检查存储FC端口状态、速率和模式
检查交换机端口状态、Zone配置、端口速率、模式
检查服务器HBA卡物理状态和驱动加载
检查多路径状态
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 49
HUAWEI

FC SAN网络不通问题分析与处理

序号	问题	解决方案
1	阵列主机口与光纤交换机协商失败	1. 更改阵列主机端口模式为点对点或者交换机模式，再重新连接 2. 更改阵列主机端口速率为1G、2G或者4G再重新连接 3. 升级阵列版本
2	交换机zone配置问题	1. 删除原来的zone； 2. 重新创建zone，保证阵列主机口和业务服务器的HBA卡在一个zone里；
3	HBA卡驱动问题	1. 卸载原有的HBA卡驱动 2. 重新安装新的HBA卡驱动
4	硬件故障	采用替换法确定故障点，是光模块还是光纤还是HBA卡故障，确定后更换。

 Page 50



FC SAN网络不通问题分析与处理

序号	问题	解决方案
5	存储单元端FC误码率过高	登录管理界面查看光纤端口误码情况，如果误码率持续增长，表示误码率过高，如果误码持续增长，属于非正常情况，应该从以下几个方面排除误码： 1）查看存储侧是否有光模块告警信息。 2）更换光纤线。 3）更换主机端口。

 Page 51 

- 针对存储单元FC误码率过高问题

1. 查看存储侧是否有光模块告警信息：是否有告警信息，存储侧光模块工作不正常，会出现大量误码，应及时更换存储对应端口光模块或 与之连接的主机侧光模块。
2. 更换光纤线：某些情况下，光纤线弯曲过大，会导致光纤内部断裂；光纤出口有沾污会导致接收光或发射光信号质量下降，这些都有可能导致数据传输中产生误码。
3. 更换主机端口：某些情况下，主机HBA卡与光纤线接触不好，可能造成光信号传输质量不好；另外，主机侧光模块工作不正常也有可能制造误码。

IP SAN网络不通问题诊断思路

序号	问题	解决方案
1	检查速度和双工模式是否为auto	**Windows平台：** 打开网络连接，本地连接状态，查看连接速度； 打开设备管理器，网卡属性，查看双工模式。 **Linux平台：** 使用命令ethtool eth0 查看连接速度和双工模式。
2	检查虚拟交换机配置	1. 确保虚拟交换机使用的物理网卡和存储业务网络连通 2. 确保虚拟机网卡选择了正确的虚拟网络
3	检查iSCSI主机端口配置	存储业务IP如果和主机IP地址不在同一子网，必须设置网关，确保路由可达。



案例1- S5600T控制器ECC复位后无法恢复上电，对端控制器离线

- 故障现象
  - S5600T，B控制器ECC错误异常复位后，无法上电成功，同时A控制器状态为“离线”。
    系统单控运行，且运行的控制器为“离线”状态。另外一个控制器恢复上电失败。
- 原因分析
  - 控制器B的内存条有异常，导致产生大量ECC错误，最终引起控制器复位，内存被屏蔽。
    由于控制器B复位后因存被屏蔽，其有效内存容量比控制器A（主控）小，控制器B恢复上电失败。
    控制器B在上电过程中，会先假定自己是主控，在上电失败时，会将对端置为“离线”，导致正在运行的真正主控（控制器A）状态变为“离线”。

 Page 53 

1. 查看复位的控制器的复位原因是ECC复位（看对应控制器故障时间点后的第一份messages日志）：
   [2012-11-21 16:04:17][2169][8001500003fa007d][INFO][Ecc reboot 1 times][OS_Proc_Init][OS_DEBUG]
2. 登录ISM查看，确认未恢复的控制器状态为“离线”。
3. 确认阵列版本是V100R002C00SPC010、V100R002C00SPC011、V100R002C00SPC012三者之一。

案例1- S5600T控制器ECC复位后无法恢复上电，对端控制器离线

- 处理过程

  1．停止主机业务。

  2．更换备件控制器。

  3．将阵列版本升级到最新的V100R002版本或V100R005版本。

  注意：如果备件控制器与当前运行控制器的软件版本不同，在版本同步过程中主控制器也会重启，如果此时未停止业务，将导致业务中断。

  检查阵列及业务状态，如果双控状态正常且版本较V100R002C00SPC012高，业务正常，则问题解决。



案例2-网线断开导致映射的LUN文件系统错误

- 故障现象
  - 应用服务器的操作系统为Linux
  - 在业务运行过程中存储设备与应用服务器之间的网线断开，且在应用服务器侧映射的LUN文件系统错误(如文件系统只读、不能正常mount等)。
- 原因分析
  - 网线损坏导致映射的LUN文件系统错误。
- 处理过程
  1. 更换网线并确保网络恢复正常。
  2. 以ext2文件系统为例，服务器上运行fsck.ext2 -a命令修复文件系统。



案例3- LUN属性错误导致对其同时读写大文件时读性能低下

- 故障现象
  - 对一个LUN同时读写两个大文件时，读性能低下，应用服务器中读文件拷贝进度剩余时间不断增加或没有显示。
- 原因分析

  1. 在ISM导航树上，展开存储设备，然后选择“存储资源 > LUN”，在右侧的信息展示区下侧查看故障的LUN的属性。

  2. 确认该LUN正在后台格式化或写策略设置为透写。

  由此得出结论：

  LUN正在后台格式化或写策略设置错误，导致对其同时读写大文件时读性能低下。



案例3- LUN属性错误导致对其同时读写大文件时读性能低下

- 处理过程
  - 等待该LUN后台格式化完成。
    登录ISM，展开存储设备，然后选择“存储资源 > LUN”，在右侧的信息展示区下侧查看故障的LUN的属性。当“运行状态”显示为“在线”时可以判断已经格式化完成。
    设置LUN写策略为回写。
    在右侧信息展示区，勾选需要修改的LUN，然后选择“修改”，在弹出的“修改LUN”对话框中，设置其写策略。

 Page 57



- 建议采用以下措施规避该故障：
  - 在同时读写大文件时，确认LUN后台格式化已完成。
  - LUN的写策略设置为回写。
  - 在LUN的后台格式化过程中或者属性设置为透写模式时，存储设备处理性能会有部分下降。

案例4-阵列CHAP信息修改，导致主机重启后无法与阵列恢复连接

- 故障现象
  - 阵列上创建CHAP用户后添加给主机启动器，并在主机和阵列侧分别启用CHAP认证功能后建立iSCSI连接。修改阵列的CHAP用户名称，更换之前的CHAP用户，重启主机后不能与阵列恢复iSCSI连接，ISM会显示启动器为“未连接”状态。
- 原因分析
  - 应用服务器重新启动且与阵列侧恢复iSCSI链路的过程中，应用服务器侧使用原有的CHAP用户信息，但阵列侧使用的是更改后的CHAP用户信息。两端CHAP用户信息不一致，导致CHAP鉴权失败，无法恢复链路。
- 处理过程
  - 手动设置主机侧CHAP信息，然后重新建立连接。





目录

 Page 59



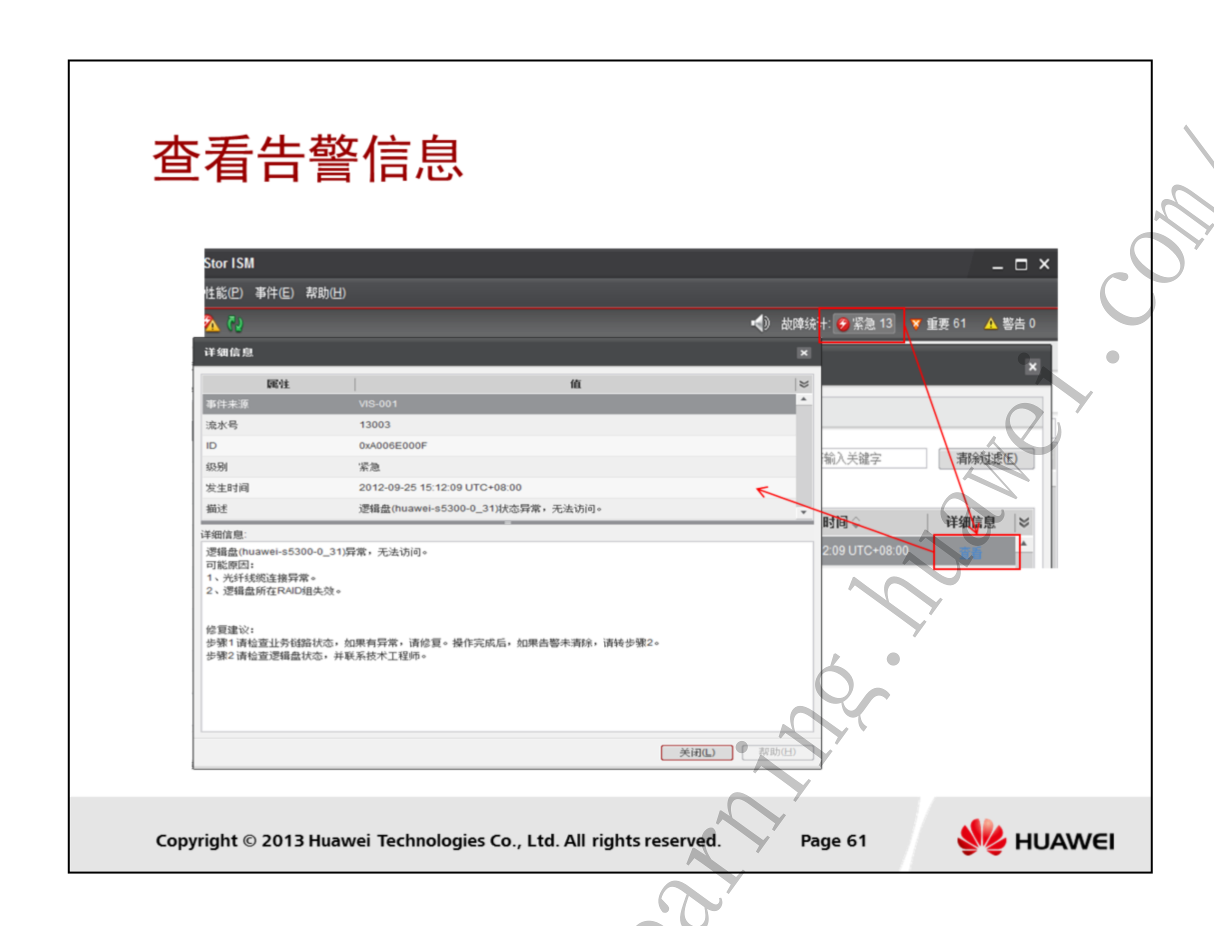
查看硬件运行状态

不正常

正常

 Page 60 HUAWEI

在ISM管理界面下的设备信息页面可以直观的查看各个硬件模块的运行状态，以上图为例：被红框圈定的电源模块的健康状态为故障。



在ISM管理界面的右上方可以看到各个告警级别的告警数量，直接点击相应的故障级别图标可以打开事件管理界面，在界面中可以看到所有详细告警信息。



在系统配置界面的导入导出部分可以对配置文件，运行数据和系统日志进行导出。

虚拟存储网关故障诊断流程

- 导出系统日志
- 登录ISM管理软件
- 根据事件处理建议处理故障
- 收集故障信息



 Page 63 HUAWEI

- 导出系统日志

  在系统出现问题时，需要尽快保存日志信息，以便对问题进行确认和定位原因。

  在寻求华为技术有限公司技术支持时，请反馈导出的系统日志。

- 登录ISM管理软件

  登录ISM管理软件，可以查询VIS6600T的运行状态、是否有告警产生。

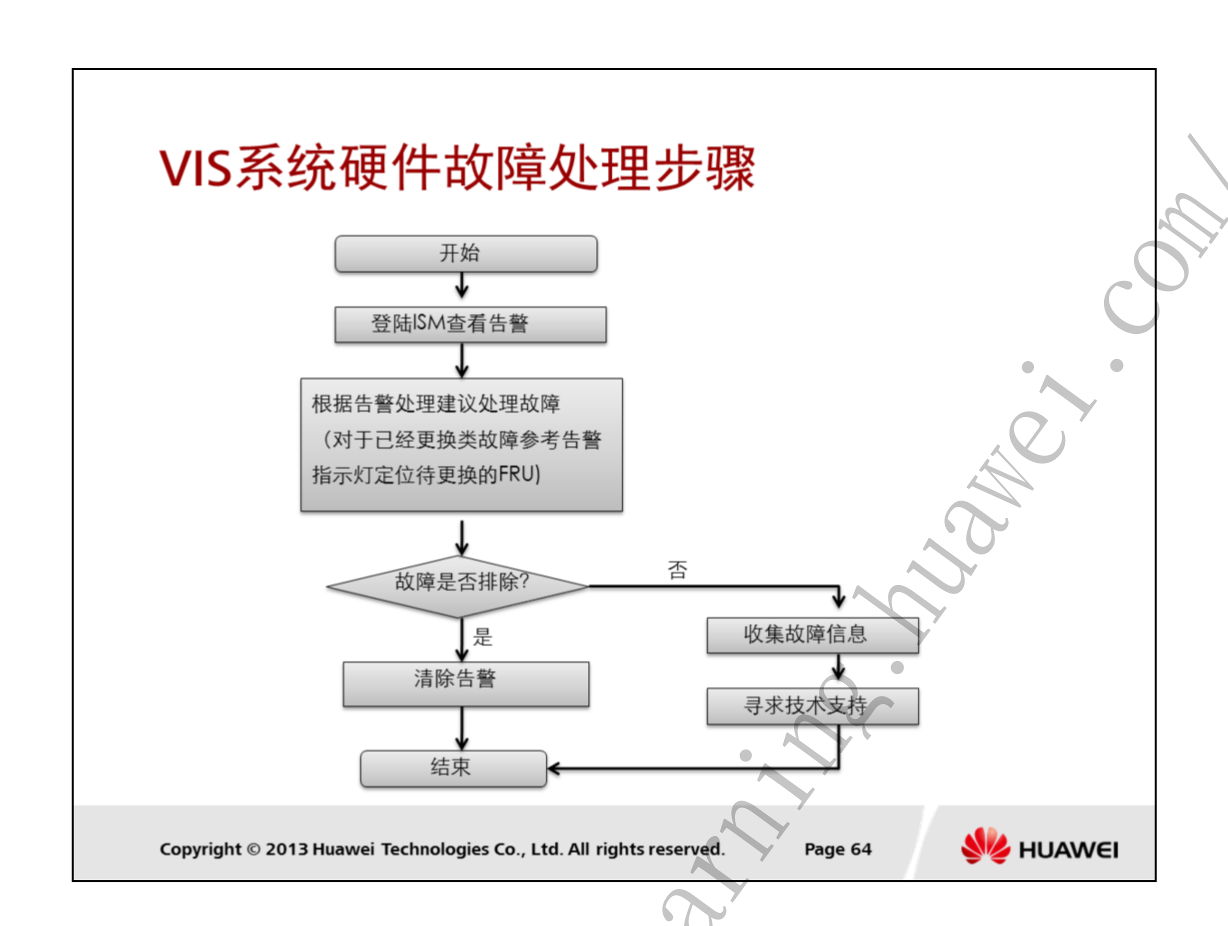
- 根据事件处理建议处理故障

  如果在ISM管理界面中查询到事件，则根据事件处理建议处理事件。

  定位故障是从众多可能原因中找出单一原因的过程，通过分析、比较各种可能的故障原因，不断排除不可能因素，最终确定故障发生的具体原因。

- 收集故障信息

  在寻求华为技术支持时，请反馈收集的故障信息。


VIS系统硬件故障处理步骤
开始
登陆ISM查看告警
根据告警处理建议处理故障
（对于已经更换类故障参考告警
指示灯定位待更换的FRU)
故障是否排除?
否
是
收集故障信息
清除告警
寻求技术支持
结束
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 64
HUAWEI

虚拟化存储网关典型故障

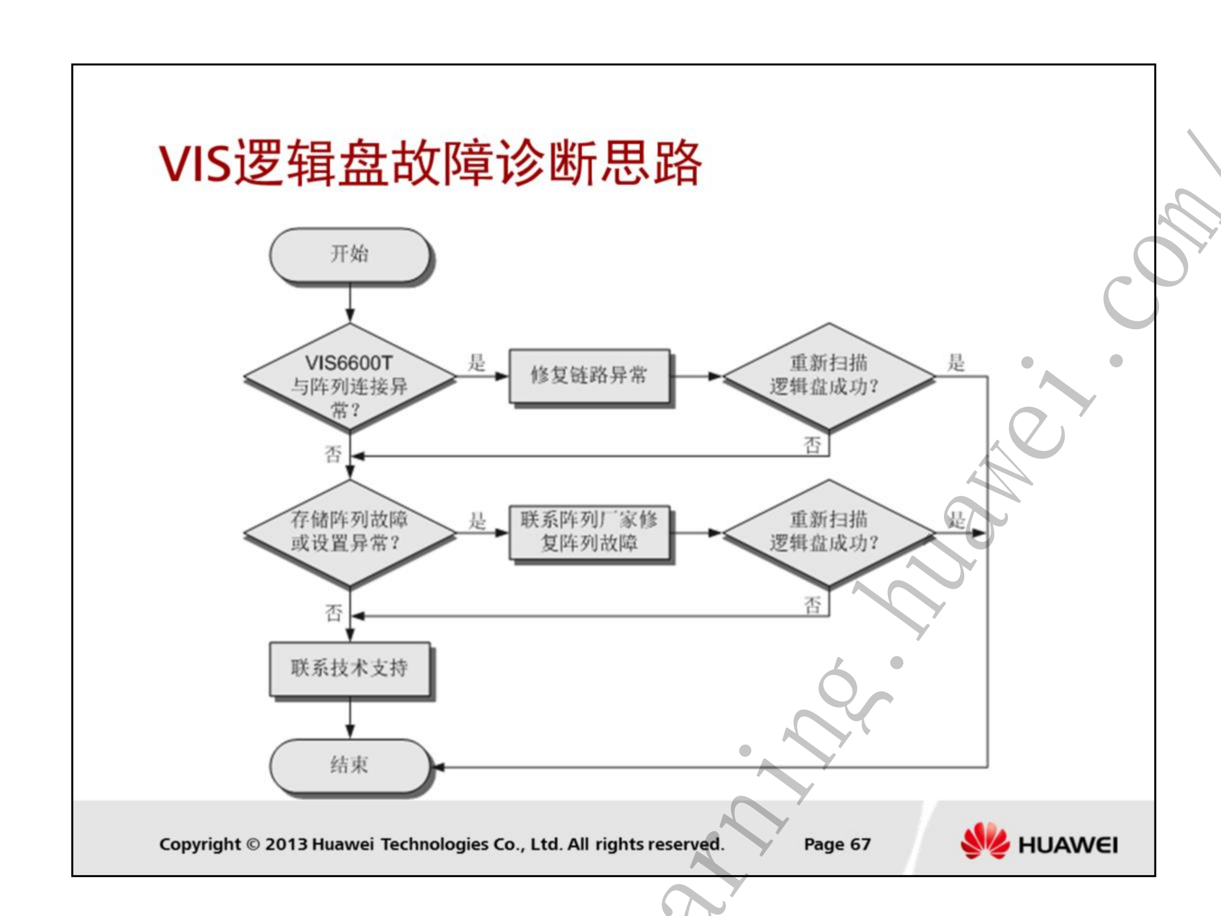
- VIS逻辑盘故障
- VIS仲裁盘故障
- 主机卷异常
- 镜像卷关系异常
- 复制卷复制异常
- 日志卷错误
- 节点上电失败
- FC链路连接异常



VIS基本概念介绍

- 元数据盘（私有盘）：元数据盘中主要存放VIS中的逻辑盘组和卷的配置信息。VIS默认配置两块元数据盘，此两块元数据盘是互为冗余的。一旦两个元数据盘全部故障，VIS的所有磁盘组和卷信息都丢失，业务也会中断。
- 仲裁盘（I/O fencing盘）：用于VIS集群成员关系变化时，确定成员角色的硬盘; 在VIS启动时，已配置的仲裁盘也必须在位。
- 逻辑盘组：多个逻辑盘组合的一个对象，是虚拟化存储管理的最小资源池; 对应VxVM中的DG。
- 卷：硬盘虚拟化管理的逻辑单元，是提供给主机应用的基本对象；对应VxVM中的LV。
- 逻辑盘：虚拟化存储管理的最小资源。对应VxVM的subdisk，对应阵列的一个LUN。

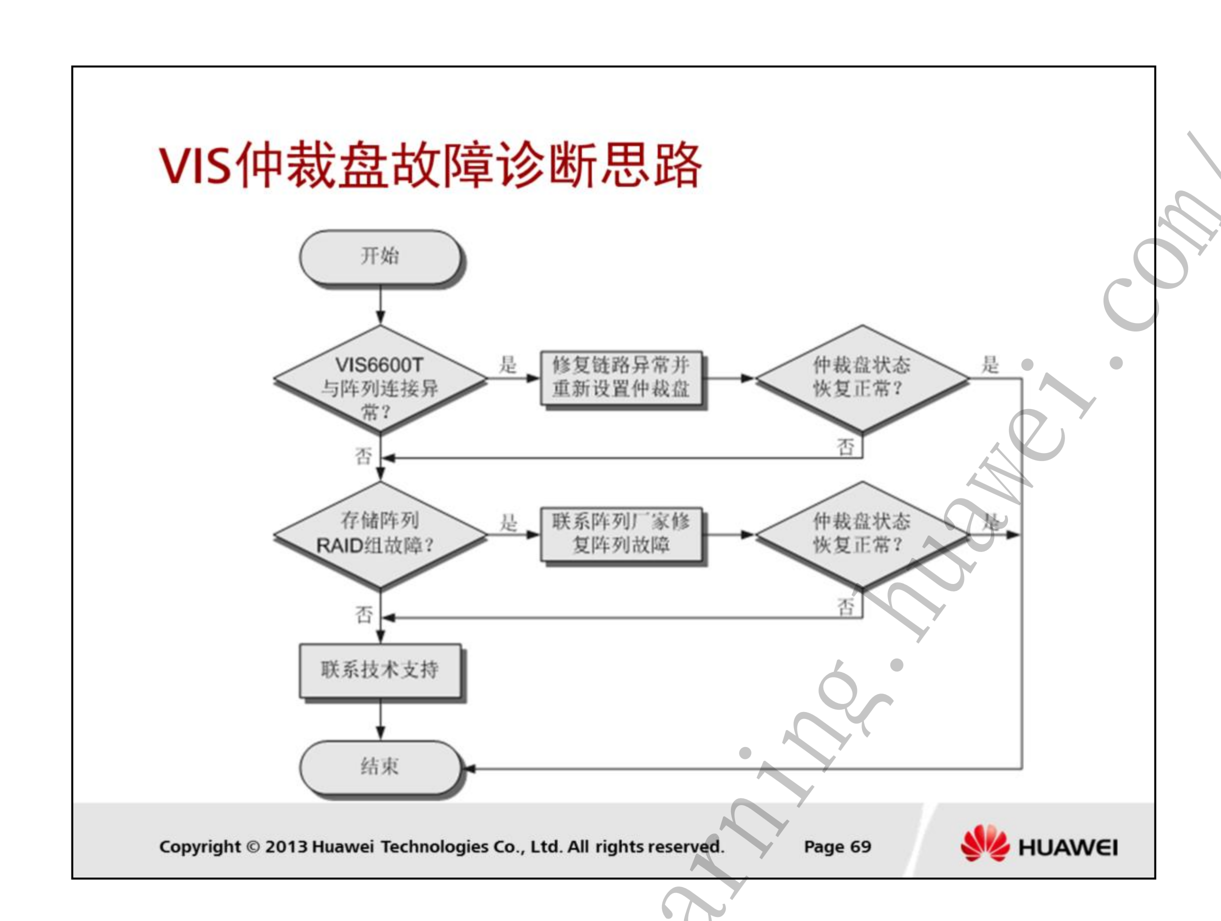



VIS逻辑盘故障诊断思路
开始
VIS6600T
与阵列连接异
常？
是
修复链路异常
重新扫描
逻辑盘成功？
是
否
否
存储阵列故障
或设置异常？
是
联系阵列厂家修
复阵列故障
重新扫描
逻辑盘成功？
是
否
否
联系技术支持
结束
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 67
HUAWEI

VIS逻辑盘故障处理方法

序号	问题	解决方案
1	VIS6600T与存储阵列的物理连接出现异常	1.确认SFP光模块与FC接口接触良好。 2.确认SFP光模块的扣环已经扣好。 3.确认光纤连接端的卡口已经卡入了SFP光模块中。 4.确认光纤的线序正确。 5.登录ISM，检查是否可以扫描到逻辑盘。 是 => 故障排除。 否 => 请进行原因2的排查。
2	存储阵列故障或阵列设置异常	1.请保持故障环境并联系存储阵列厂商，检查并修复阵列故障和异常设置。 2.登录ISM，检查是否可以扫描到逻辑盘。 是 => 故障排除。 否 => 请保持故障环境并联系技术支持工程师进行处理。




VIS仲裁盘故障诊断思路
开始
VIS6600T
与阵列连接异
常？
是
修复链路异常并
重新设置仲裁盘
仲裁盘状态
恢复正常？
是
否
否
存储阵列
RAID组故障？
是
联系阵列厂家修
复阵列故障
仲裁盘状态
恢复正常？
是
否
否
联系技术支持
结束
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 69
HUAWEI

VIS仲裁盘故障处理方法

序号	问题	解决方案
1	VIS6600T与仲裁盘所属的阵列连接异常	1.检查VIS6600T与仲裁盘所属的阵列的FC主机端口Link指示灯状态和速率指示灯状态是否正常。 2.重新连接VIS6600T与仲裁盘所属的阵列后，在ISM管理界面上重新设置仲裁盘。 3.在ISM管理界面中，检查仲裁盘的告警是否清除。

 Page 70 

- 针对VIS6600T与仲裁盘所属的阵列连接异常

  1.检查VIS6600T与仲裁盘所属的阵列的FC主机端口Link指示灯状态和速率指示灯状态是否正常。

  是 => 请进行原因2的排查。

  否 => 请继续执行步骤1.2。

  2.重新连接VIS6600T与仲裁盘所属的阵列后，在ISM管理界面上重新设置仲裁盘。

  在导航树上选择“所有设备 > VIS6600T > 存储资源 > 逻辑盘”。

  在右侧信息展示区单击“扫描逻辑盘”，重新扫描逻辑盘。

  扫描逻辑盘完成后，在右侧信息展示区单击“设置仲裁盘”，系统弹出“设置仲裁盘”对话框。

  完成设置仲裁盘的操作，单击“确定”。

  3.在ISM管理界面中，检查仲裁盘的告警是否清除。

  是 => 故障排除。否 => 请进行原因2的排查。

VIS仲裁盘故障处理方法

序号	问题	解决方案
2	仲裁盘所属阵列的RAID组故障	1.请保持故障环境并联系存储阵列厂商的技术支持工程师，修复存储阵列的RAID组故障。 2.在ISM管理界面上重新设置仲裁盘。 3.在ISM管理界面中，检查仲裁盘的告警是否清除。

 Page 71



- 针对仲裁盘所属阵列的RAID组故障

  1.请保持故障环境并联系存储阵列厂商的技术支持工程师，修复存储阵列的RAID组故障。

  2.在ISM管理界面上重新设置仲裁盘。

  在导航树上选择“所有设备 > VIS6600T > 存储资源 > 逻辑盘”。

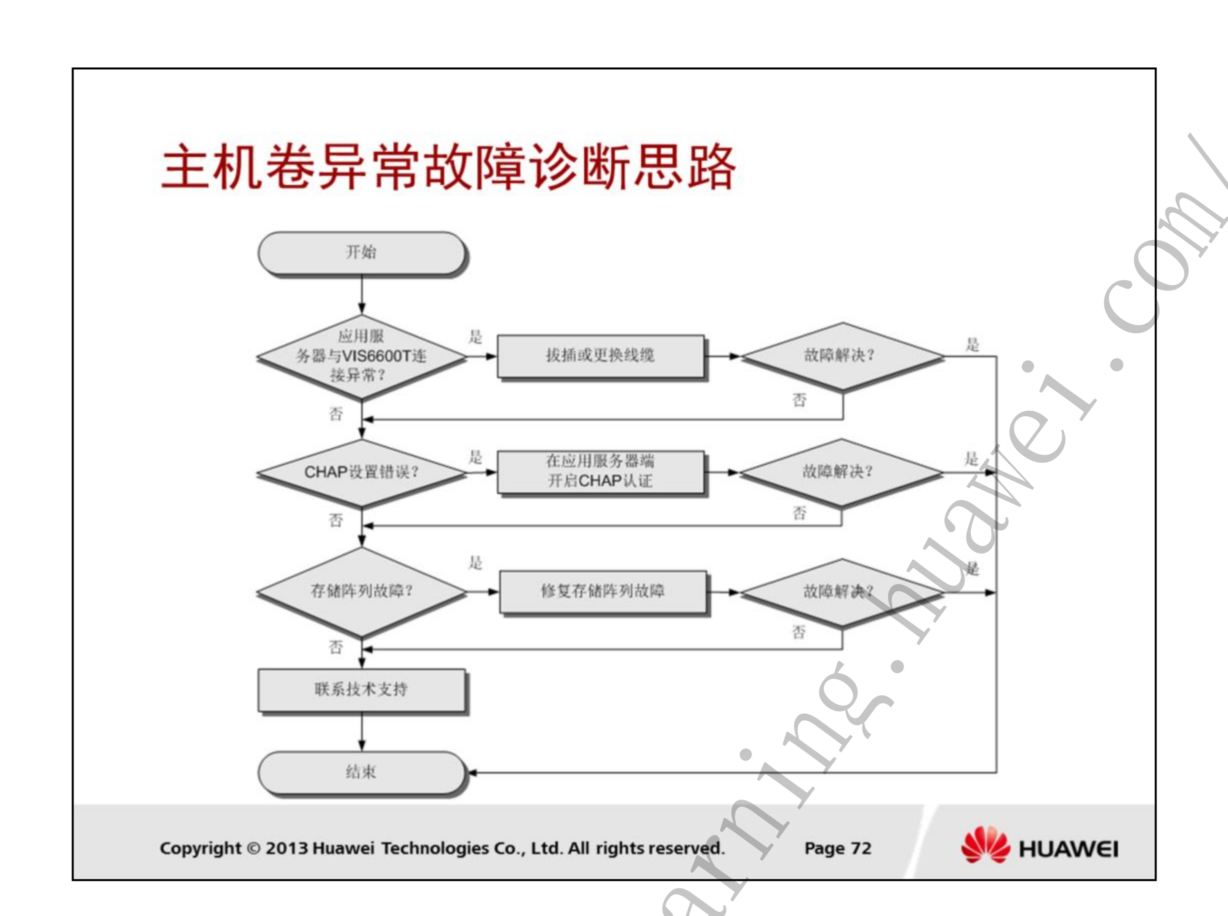
  在右侧信息展示区单击“扫描逻辑盘”，重新扫描逻辑盘。

  扫描逻辑盘完成后，在右侧信息展示区单击“设置仲裁盘”，系统弹出“设置仲裁盘”对话框。

  完成设置仲裁盘的操作，单击“确定”。

  3.在ISM管理界面中，检查仲裁盘的告警是否清除。

  是 => 故障排除。否 => 请保持故障环境并联系技术支持工程师进行处理。


主机卷异常故障诊断思路
开始
应用服务器与VIS6600T连接异常？
是
拔插或更换线缆
故障解决？
是
否
否
CHAP设置错误？
是
在应用服务器端开启CHAP认证
故障解决？
是
否
否
存储阵列故障？
是
修复存储阵列故障
故障解决？
是
否
否
联系技术支持
结束
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 72
HUAWEI

主机卷异常故障处理方法

序号	问题	解决方案
1	应用服务器与VIS6600T的连接异常	1.确认应用服务器与VIS6600T的连接方式。 2.应用服务器与VIS6600T使用光纤连接。 a.在设备侧检查VIS6600T与应用服务器连接的FC主机端口link指示灯是否熄灭。 b.拔插光纤或更换光纤。 c.操作结束后，在设备侧检查VIS6600T与应用服务器连接的FC主机端口link指示灯是否熄灭。 d.在应用服务器端重新扫描硬盘，查看扫描是否成功。 3.应用服务器与VIS6600T使用网线连接。 a.在设备侧检查VIS6600T与应用服务器连接的iSCSI主机端口link指示灯是否熄灭。 b.拔插网线或更换网线。 c.操作结束后，在设备侧检查VIS6600T与应用服务器连接的iSCSI主机端口link指示灯是否熄灭。 d.在应用服务器端重新扫描硬盘，查看扫描是否成功。



- 针对应用服务器与VIS6600T的连接异常

1.确认应用服务器与VIS6600T的连接方式。

光纤 => 请继续执行步骤2。

网线 => 请继续执行步骤3。

2.应用服务器与VIS6600T使用光纤连接。

a.在设备侧检查VIS6600T与应用服务器连接的FC主机端口link指示灯是否熄灭。
是 => 请继续执行2.b。否 => 请进行原因2的排查。

b.拔插光纤或更换光纤。
c.操作结束后，在设备侧检查VIS6600T与应用服务器连接的FC主机端口link指示灯是否熄灭。
否 => 请继续执行2.d。是 => 请进行原因2的排查。

d.在应用服务器端重新扫描硬盘，查看扫描是否成功。
是 => 故障排除。否 => 请进行原因2的排查。

3.应用服务器与VIS6600T使用网线连接。

a.在设备侧检查VIS6600T与应用服务器连接的iSCSI主机端口link指示灯是否熄灭。
是 => 请继续执行3.b。否 => 请进行原因2的排查。

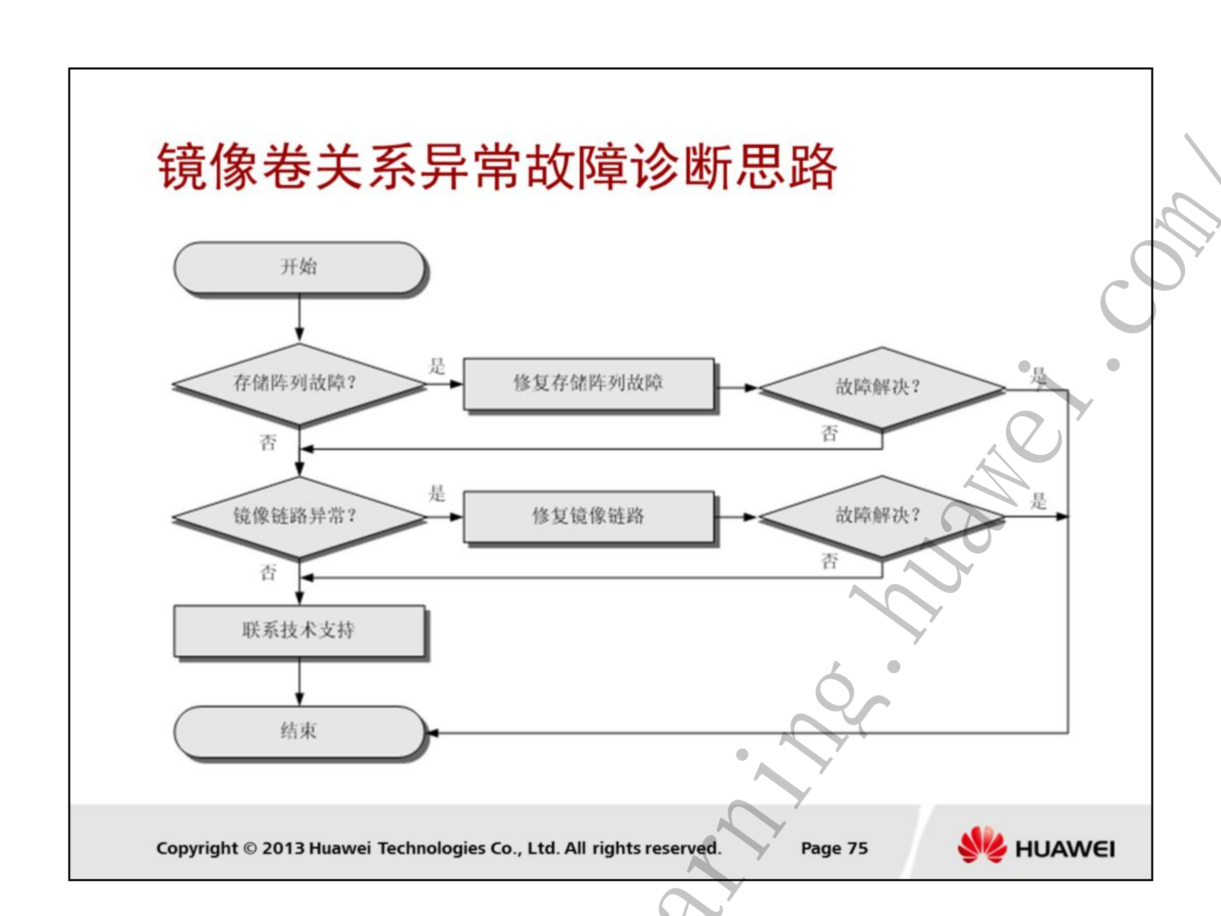
b.拔插网线或更换网线。
c.操作结束后，在设备侧检查VIS6600T与应用服务器连接的iSCSI主机端口link指示灯是否熄灭。
否 => 请继续执行3.d。是 => 请进行原因2的排查。

d.在应用服务器端重新扫描硬盘，查看扫描是否成功。
是 => 故障排除。否 => 请进行原因2的排查。

主机卷异常故障处理方法

序号	问题	解决方案
2	CHAP设置有误	1.如果在VIS6600T开启了CHAP认证，检查在应用服务器端是否开启CHAP认证。 是 => 请进行原因3的排查。 否 => 在应用服务器端开启CHAP认证，详细信息请参见应用服务器相关资料。开启CHAP认证后，请继续执行步骤2。 2.在应用服务器端重新扫描硬盘，查看扫描是否成功。 是 => 故障排除。否 => 请进行原因3的排查。
3	存储阵列故障	1.请检查存储阵列是否故障。 是 => 请根据存储阵列的相关说明修复故障。存储阵列的故障修复后，进入ISM管理界面重新执行扫描逻辑盘等操作。 否 => 请保持故障环境并联系技术支持工程师进行处理。 2.在应用服务器端重新扫描硬盘，查看扫描是否成功。 是 => 故障排除。 否 => 请保持故障环境并联系技术支持工程师进行处理。




镜像卷关系异常故障诊断思路
开始
存储阵列故障？
是
修复存储阵列故障
故障解决？
是
否
否
镜像链路异常？
是
修复镜像链路
故障解决？
是
否
否
联系技术支持
结束
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 75
HUAWEI

镜像卷关系异常故障处理方法

序号	问题	解决方案
1	存储阵列故障	1.请检查存储阵列是否发生故障。 2.请根据存储阵列的相关说明修复故障，检查故障是否清除。 3.登录ISM管理界面，重新执行扫描逻辑盘和恢复镜像的操作。在ISM管理界面上查看镜像关系是否正常。
2	镜像链路异常	1.请检查VIS6600T与存储阵列的连接状态是否为异常。 2.请修复VIS6600T与存储阵列的连接，检查连接是否修复。 3.登录ISM管理界面，重新执行扫描逻辑盘和恢复镜像的操作。在ISM管理界面上查看镜像关系是否正常。



- 针对存储阵列故障

1.请检查存储阵列是否发生故障。

是 => 请继续执行步骤2。否 => 请进行原因2的排查。

2.请根据存储阵列的相关说明修复故障，检查故障是否清除。

是 => 请继续执行步骤3。否 => 请保持故障环境并联系存储阵列厂商技术支持工程师进行处理。

3.登录ISM管理界面，重新执行扫描逻辑盘和恢复镜像的操作。在ISM管理界面上查看镜像关系是否正常。

是 => 故障排除。否 => 请保持故障环境并联系技术支持工程师进行处理。

- 针对镜像链路异常

1.请检查存储阵列是否发生故障。

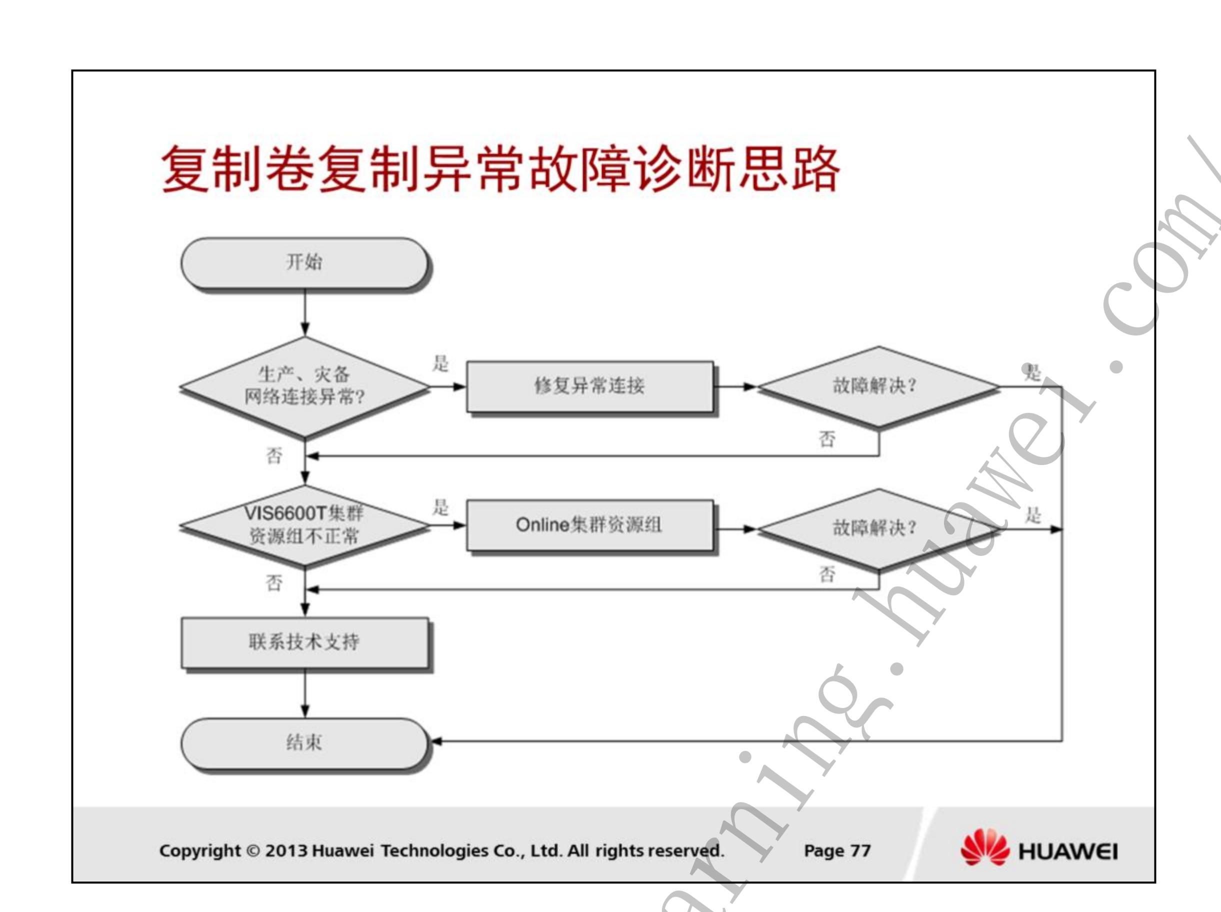
是 => 请继续执行步骤2。否 => 请进行原因2的排查。

2.请根据存储阵列的相关说明修复故障，检查故障是否清除。

是 => 请继续执行步骤3。否 => 请保持故障环境并联系存储阵列厂商技术支持工程师进行处理。

3.登录ISM管理界面，重新执行扫描逻辑盘和恢复镜像的操作。在ISM管理界面上查看镜像关系是否正常。

是 => 故障排除。否 => 请保持故障环境并联系技术支持工程师进行处理。


复制卷复制异常故障诊断思路
开始
生产、灾备
网络连接异常?
是
修复异常连接
故障解决？
是
否
否
VIS6600T集群
资源组不正常
是
Online集群资源组
故障解决？
是
否
否
联系技术支持
结束
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 77
HUAWEI

复制卷复制异常故障处理方法

序号	问题	解决方案
1	生产中心、灾备中心网络连接不畅通	1.登录ISM，检查端口的状态。 2.在ISM导航树中选择“所有设备 > VIS6600T > 逻辑盘组”，选择创建了复制一致性组的逻辑盘组，选择“复制一致性组”页签，获取生产中心的虚拟IP地址。 3.使用相同方法获取灾备中心的虚拟IP地址。 4.登录CLI（通过Putty等工具），检查生产中心与灾备中心的连接状态。 5.重新启动复制卷复制。 6.检查复制卷复制异常或中断的情况是否消失。

 Page 78



1.登录ISM，检查端口的状态。
通过执行发现设备操作，分别发现生产中心的VIS6600T和灾备中心的VIS6600T。进入“事件管理”对话框，查看是否有端口的相关告警。如果有，请根据修复建议进行修复。

2.在ISM导航树中选择“所有设备 > VIS6600T > 逻辑盘组”，选择创建了复制一致性组的逻辑盘组，选择“复制一致性组”页签，获取生产中心的虚拟IP地址。
3.使用相同方法获取灾备中心的虚拟IP地址。
4.登录CLI（通过Putty等工具），检查生产中心与灾备中心的连接状态。
a.登录生产中心的VIS6600T主节点的CLI界面，使用命令**vxping** *XXX.XXX.XXX.XXX*（XXX.XXX.XXX.XXX代表灾备中心的虚拟IP地址）查看生产中心与灾备中心连接是否畅通。
b.登录灾备中心的VIS6600T主节点的CLI界面，使用命令**vxping** *XXX.XXX.XXX.XXX*（XXX.XXX.XXX.XXX代表生产中心的虚拟IP地址）查看生产中心与灾备中心连接是否畅通。
说明：

在CLI界面输入**vxdctl -c mode**可以辨别主控制器。

```
CLI:admin>vxdctl -c mode
mode: enabled: cluster active - MASTER
master: VIS6600T_6695_1
```

如果生产中心与灾备中心网络不通，请检查生产中心与灾备中心连接的iSCSI主机端口是否在同一网段，如果不在同一网段，请为iSCSI主机端口配置相应的路由信息。配置路由的详细信息请参见《OceanStor VIS6600T 虚拟智能存储系统 配置指南》。此外需要检查网络中的防火墙是否屏蔽了复制的端口号。

5.重新启动复制卷复制。
6.检查复制卷复制异常或中断的情况是否消失。
是 => 故障解决。否 => 进行原因2的排查。

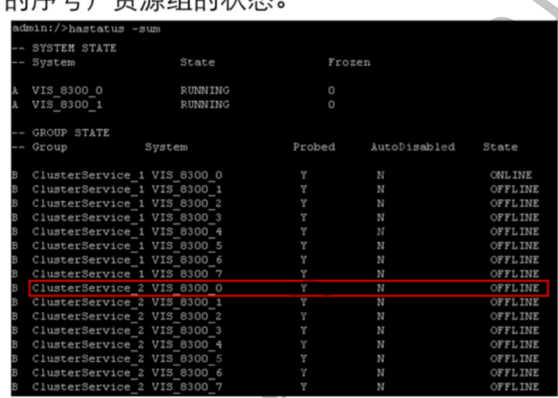
复制卷复制异常故障处理方法

原因2：VIS6600T集群资源组不正常

解决方案

1.登录CLI（通过Putty等工具），执行命令**hastatus –sum**查看主节点的ClusterService_X（X表示虚拟IP的序号）资源组的状态。

- “OFFLINE” => 请继续执行步骤2。
- “FAULT” => 请继续执行步骤3。
- “ONLINE” => 请保持故障环境并联系支持工程师。



```
admin:/>hastatus -sum
-- SYSTEM STATE
-- System               State              Frozen

A  VIS_8300_0           RUNNING            0
A  VIS_8300_1           RUNNING            0

-- GROUP STATE
-- Group            System          Probed     AutoDisabled    State

B  ClusterService_1 VIS_8300_0        Y          N               ONLINE
B  ClusterService_1 VIS_8300_1        Y          N               OFFLINE
B  ClusterService_1 VIS_8300_2        Y          N               OFFLINE
B  ClusterService_1 VIS_8300_3        Y          N               OFFLINE
B  ClusterService_1 VIS_8300_4        Y          N               OFFLINE
B  ClusterService_1 VIS_8300_5        Y          N               OFFLINE
B  ClusterService_1 VIS_8300_6        Y          N               OFFLINE
B  ClusterService_1 VIS_8300_7        Y          N               OFFLINE
B  ClusterService_2 VIS_8300_0        Y          N               OFFLINE
B  ClusterService_2 VIS_8300_1        Y          N               OFFLINE
B  ClusterService_2 VIS_8300_2        Y          N               OFFLINE
B  ClusterService_2 VIS_8300_3        Y          N               OFFLINE
B  ClusterService_2 VIS_8300_4        Y          N               OFFLINE
B  ClusterService_2 VIS_8300_5        Y          N               OFFLINE
B  ClusterService_2 VIS_8300_6        Y          N               OFFLINE
B  ClusterService_2 VIS_8300_7        Y          N               OFFLINE
```

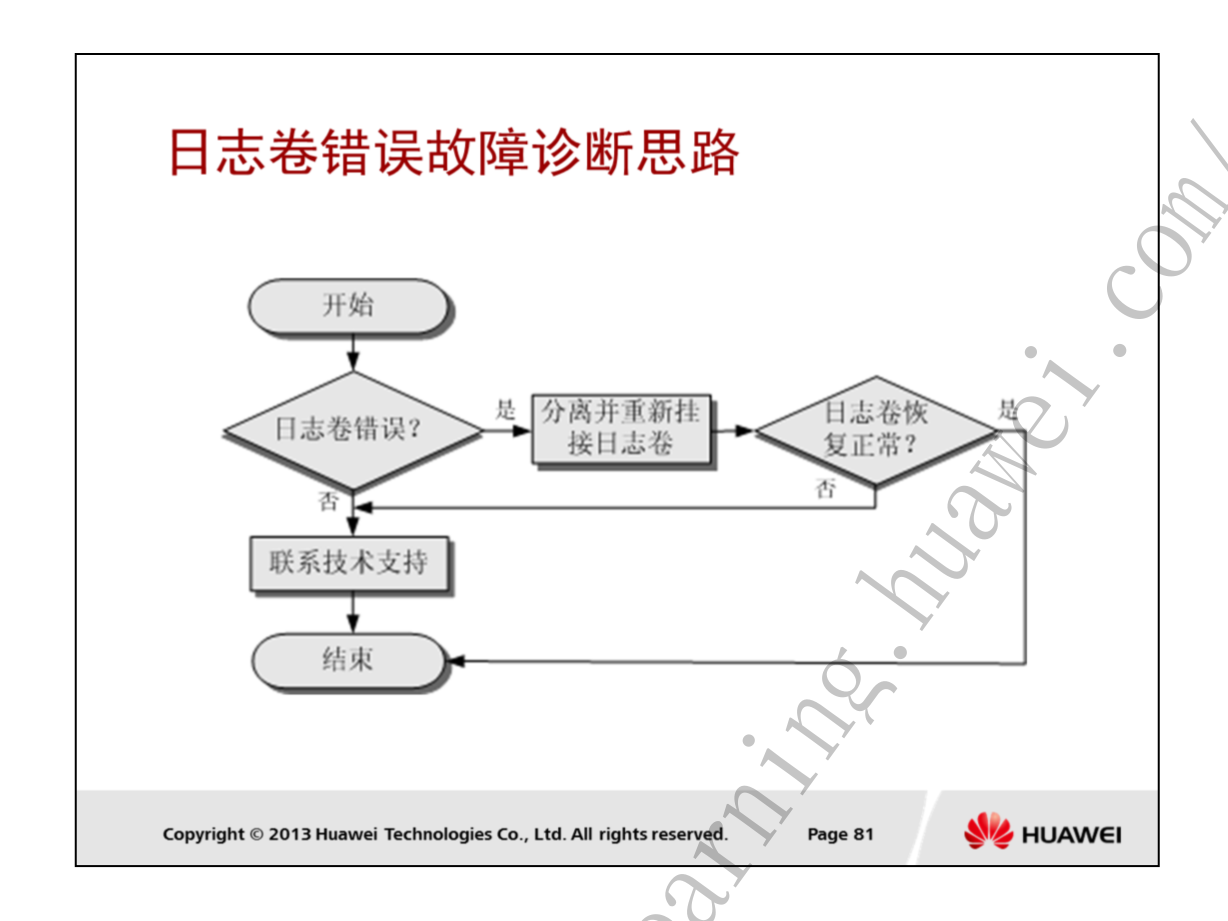



复制卷复制异常故障处理方法

原因2：VIS6600T集群资源组不正常

解决方案
2.请执行如下命令： hagrp -online ClusterService_X -sys VIS6600T_6695_1（在VIS6600T_6695_1上online资源）。 请继续执行步骤4。 3.请依次执行命令： hagrp -clear ClusterService_X -sys VIS6600T_6695_1（清除VIS6600T_6695_1的fault标记）。 hagrp -online ClusterService_X -sys VIS6600T_6695_1（在VIS6600T_6695_1上online资源）。 请继续执行步骤4。 4.执行命令hastatus –sum，查看主节点的ClusterService_X（X表示虚拟IP的序号）资源组的状态是否为“ONLINE”。 是 => 故障解决。否 => 请继续执行步骤5。 5.请执行如下命令：vxrestartvvr -f和vxrestartvxvm -f。 6.执行命令hastatus –sum，查看主节点的ClusterService_X（X表示虚拟IP的序号）资源组的状态是否为“ONLINE”。 是 => 故障解决。否 => 请保持故障环境并联系技术支持工程师进行处理。





- 现象描述

  在配置了复制一致性组的情况下，由于日志卷错误，导致生产中心的数据无法同步到灾备中心。

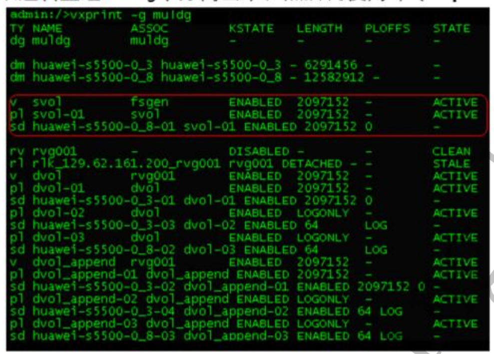
- 可能原因

  由于重新启动、逻辑盘故障修复等原因，可能会导致日志卷的头部存在部分脏数据或日志卷故障，从而导致RVG进入pass-thru模式，不能进行同步。

日志卷错误故障处理方法

原因1：由于重新启动、逻辑盘故障修复等原因，可能会导致日志卷的头部存在部分脏数据或日志卷故障，从而导致RVG进入pass-thru模式，不能进行同步。

解决方案
1.登录CLI（通过Putty等工具），执行命令**vxvol -g** *diskgroup* **-f dis** *srl*，将日志卷从复制一致性组中分离出来。其中*diskgroup*为逻辑盘组名称，*srl*为日志卷名称。 例如：先将日志卷svol从逻辑盘组muldg中分离出来，然后再使用命令**vxprint –g** *muldg*查看。





日志卷错误故障处理方法

解决方案
2.使用命令vxvol –g diskgroup aslog rvg srl将分离出来的日志卷重新挂接回复制一致性组。其中diskgroup为逻辑盘组名称，rvg为复制一致性组名称，srl为日志卷名称。 3.执行开始同步操作。 a.登录到ISM管理界面，在导航树上选择“逻辑盘组”节点中需要执行同步的逻辑盘组。 b.在右侧信息展示区的“复制一致性组”页签中选择需要执行开始同步操作的逻辑盘组。 c.单击“灾备应用管理”。系统弹出“灾备应用管理”对话框。 d.单击“设置复制状态”。系统弹出“设置复制状态”对话框。 e.单击“开始”。系统弹出“开始复制一致性组”对话框。 f.单击“确定”。系统提示操作成功。 4.重新启动复制业务。 5.检查日志卷错误无法开始同步的情况是否消失。 是 => 故障排除。否 => 请保持故障环境并联系技术支持工程师进行处理。



节点上电失败诊断思路

- 现象描述
  - 在菜单栏上选择“事件 > 事件管理”，在弹出的“事件管理”对话框的“故障列表”中有“节点上电失败”的告警。
- 可能原因
  - 原因1：本地软件版本与主节点软件版本不一致。
  - 原因2：本地CVM启动失败。
  - 原因3：节点个数超过License限制。



节点上电失败处理方法

解决方案
1.登录CLI（通过Putty等工具），执行**showloadfailreason**命令，显示上电失败的错误信息。 admin:/>showloadfailreason -c xx ==================================== System Load Failure Reason ==================================== ************************************** * Node ID xx * * Reason xxxxxxxxxxxxxxx * ************************************** 其中“xx”表示节点数。 如果界面显示信息如下： System Check Local Version Failure => 请直接执行步骤2~步骤3。 Local Start CVM Failure => 请直接执行步骤4~步骤8。 Node Number of License Is Not Enough => 请直接执行步骤9。



节点上电失败处理方法

解决方案

2.执行**showcontroller**查看是否存在某个节点的版本号“Release”与主节点版本号“Release”不一致。

admin:/>showcontroller

```
==============================================
          Controller Information
----------------------------------------------
Controller ID              | 0                      Controller ID              | 1
Frame ID                   | 0                      Frame ID                   | 0
Slot ID                    | 0                      Slot ID                    | 0
Controller IP              | 100.184.125.97         Controller IP              | 100.184.125.98
Master                     | No                     Master                     | Yes
Status                     | normal                 Status                     | normal
Release                    | 2.05.03.104.T.02       Release                    | 2.05.03.104.T.02
BIOS Version               | 06.03.03.T01           BIOS Version               | 06.03.03.T01
BMC Version                | 3.04T13                BMC Version                | 3.04T13
Operation Software Version | 2.05.03.104.T.02       Operation Software Version | 2.05.03.104.T.02
Master CPLD Version        | 230T01                 Master CPLD Version        | 230T01
Standby CPLD Version       | 320T02                 Standby CPLD Version       | 320T02
Location                   | Up                     Location                   | Up
Temperature Status         | Normal                 Temperature Status         | Normal
Voltage Status             | Normal                 Voltage Status             | Normal
----------------------------------------------
==============================================
```

是 => 请继续执行步骤3。否 => 请保持故障环境并联系技术支持工程师进行处理。

 Page 86 HUAWEI

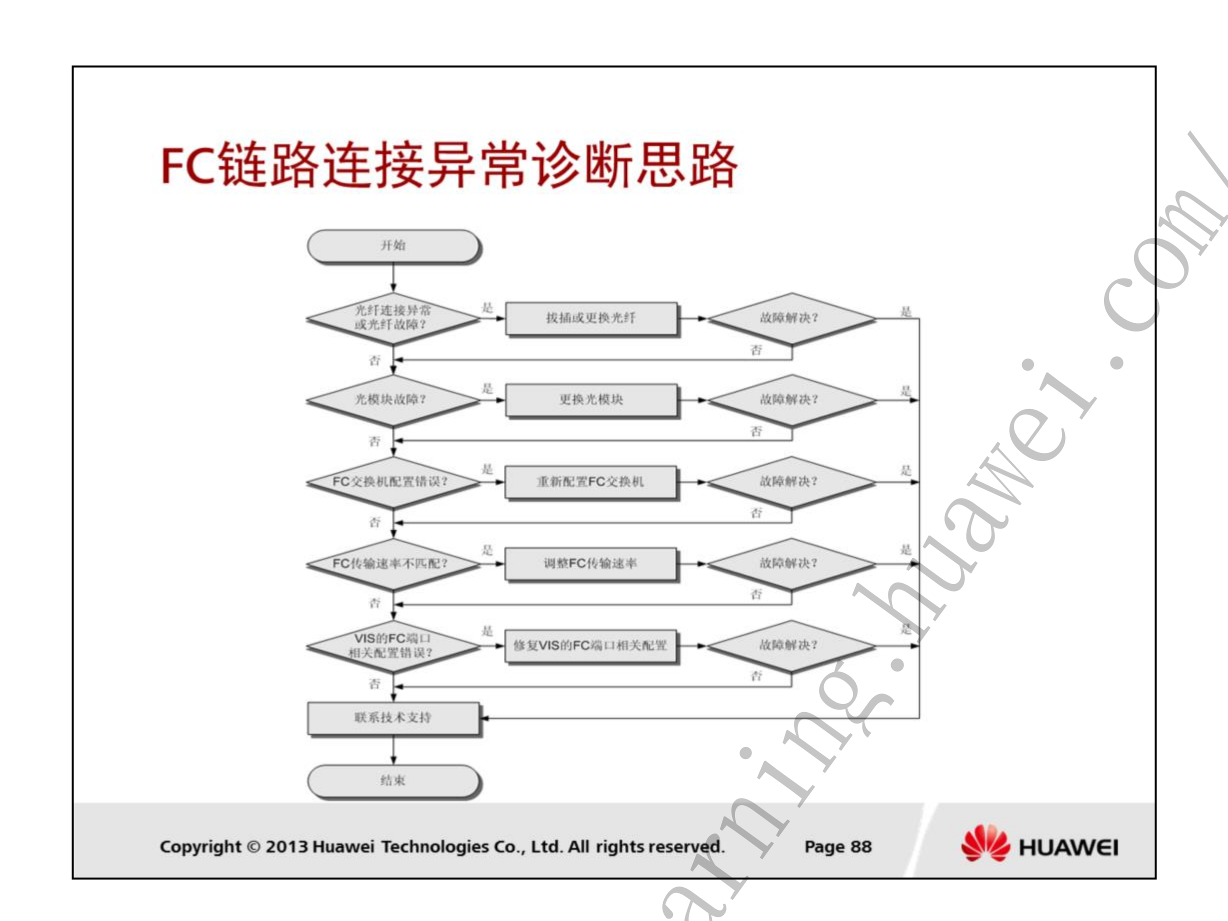
*说明：在CLI界面输入**vxdctl -c mode**可以查看主节点信息。*

admin:/>vxdctl -c mode
mode: enabled: cluster active - MASTER
master: VIS_6796_0

节点上电失败处理方法

解决方案
3.将该节点重新启动，检查该节点是否启动成功。 是 => 故障排除。否 => 请保持故障环境并联系技术支持工程师进行处理。 4.执行hastatus -sum查询集群状态是否正常。 是 => 请继续执行步骤5。否 => 请保持故障环境并联系技术支持工程师进行处理。 5.在集群各个节点上执行vxappend shwfendsk。 检查仲裁盘个数是否都为3个，且3个仲裁盘的"STATUS"都为"Normal"。 是 => 请保持故障环境并联系技术支持工程师进行处理。否 => 请继续执行步骤6。 6.请在得到技术支持工程师的确认后，断开业务。 7.执行vxappend clrfen命令，清除仲裁盘配置。 8.重新启动系统，检查系统启动是否成功。 是 => 故障排除。否 => 请保持故障环境并联系技术支持工程师进行处理。 系统启动完成后，需要重新配置仲裁盘。详细信息请参见《OceanStor VIS6600T 虚拟智能存储系统 安装指南》初始配置章节。 9.执行showlicenseresource查询License支持的节点数。 检查实际节点数是否超过licence支持的节点数。 是 => 请购买新的license。否 => 请保持故障环境并联系技术支持工程师进行处理。




FC链路连接异常诊断思路
开始
光纤连接异常或光纤故障？
是
拔插或更换光纤
故障解决？
是
否
否
光模块故障？
是
更换光模块
故障解决？
是
否
否
FC交换机配置错误？
是
重新配置FC交换机
故障解决？
是
否
否
FC传输速率不匹配？
是
调整FC传输速率
故障解决？
是
否
否
VIS的FC端口相关配置错误？
是
修复VIS的FC端口相关配置
故障解决？
是
否
否
联系技术支持
结束
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 88
HUAWEI

FC链路连接异常诊断思路

序号	问题	解决方案
1	光纤连接异常或光纤出现故障	1.在设备侧检查FC主机端口link指示灯是否熄灭。 2.拔插光纤或更换光纤。 3.操作结束后，在设备侧检查FC主机端口link指示灯是否熄灭。
2	光模块出现故障	1.依次更换VIS6600T和应用服务器上连接异常的光模块。 2.操作结束后，在设备侧检查FC主机端口link指示灯是否熄灭。

 Page 89



- 针对光纤连接异常或光纤出现故障
  1. 在设备侧检查FC主机端口link指示灯是否熄灭。

     是 => 请继续执行步骤2。否 => 请进行原因2的排查。
  2. 拔插光纤或更换光纤。

     注意：拔插光纤的过程需要注意，在拔出光纤后，等待10秒，再插入光纤。
  3. 操作结束后，在设备侧检查FC主机端口link指示灯是否熄灭。

     否 => 故障排除。是 => 请进行原因2的排查。
- 针对光模块出现故障
  1. 依次更换VIS6600T和应用服务器上连接异常的光模块。
  2. 操作结束后，在设备侧检查FC主机端口link指示灯是否熄灭。

     否 => 故障排除。是 => 请进行原因3的排查。

FC链路连接异常诊断思路

序号	问题	解决方案
3	FC交换机配置错误	1.联系FC交换机厂商，检查交换机的配置及zone划分等是否正确。 2.操作结束后，在设备侧检查FC主机端口link指示灯是否熄灭。
4	应用服务器FC HBA卡速率与VIS6600T的FC主机端口速率不匹配	1.记录应用服务器中FC HBA卡设置的工作速率。 2.确认FC主机端口速率与服务器是否一致。 3.将VIS6600T的FC主机端口的速率调整为与应用服务器FC HBA卡的速率一致。 4.操作结束后，在设备侧检查FC主机端口link指示灯是否熄灭。

 Page 90



- 针对FC交换机配置错误

1. 联系FC交换机厂商，检查交换机的配置及zone划分等是否正确。
2. 操作结束后，在设备侧检查FC主机端口link指示灯是否熄灭。

   否 => 故障解决。是 => 请进行原因4的排查。

- 针对应用服务器FC HBA卡速率与VIS6600T的FC主机端口速率不匹配

1. 记录应用服务器中FC HBA卡设置的工作速率。请参见应用服务器相关资料。
2. 确认FC主机端口速率与服务器是否一致。
   - 在ISM管理界面，选择“所有设备 > VIS6600T > 设备信息”；
   - 选择“业务控制单元 > 端口”，
   - 在右边的信息展示区查看FC主机端口的“配置速率”与相应的应用服务器FC HBA卡工作速率是否相同。

     是 => 请保持故障环境并联系技术支持工程师进行处理。

     否 => 请继续执行步骤4。
3. 将VIS6600T的FC主机端口的速率调整为与应用服务器FC HBA卡的速率一致。
4. 操作结束后，在设备侧检查FC主机端口link指示灯是否熄灭。

     否 => 故障排除。是 => 请进行原因5的排查。

FC链路连接异常诊断思路

序号	问题	解决方案
5	VIS6600T的FC主机端口相关配置错误	1.确认FC端口配置。 2.请根据具体组网环境修改FC主机端口的各项配置。 3.操作结束后，在设备侧检查FC主机端口link指示灯是否熄灭。

 Page 91



- 针对VIS6600T的FC主机端口相关配置错误

1. 确认FC端口配置。
   - 登录ISM，点击“所有设备 > VIS6600T > 设备信息 > 业务控制单元 > 端口”
   - 在右侧信息展示区中选择“FC主机端口”页签；
   - 分别选择各FC主机端口，点击“修改”；
   - 系统弹出“修改FC主机端口”对话框；
   - 分别检查各FC主机端口的“配置速率”、“端口模式”的配置项是否正确。

   是 => 请保持故障环境并联系技术工程师。否 => 请继续执行步骤4。
2. 请根据具体组网环境修改FC主机端口的各项配置。
3. 操作结束后，在设备侧检查FC主机端口link指示灯是否熄灭。

   否 => 故障解决。是 => 请保持故障环境并联系技术工程师。



思考题

1. 故障诊断处理的原则流程有哪些?
2. 连接SAN存储的主机业务中断，需要如何解决?





总结

- 故障诊断原则，流程和方法
- SAN存储故障处理思路和方法
- VIS存储故障处理思路和方法





习题

- 判断题
  1. 故障诊断应先看高级别告警再看低级别告警（T of F)
- 多选题
  1. 故障定位的主要方法有？（　　）

     A.重启系统

     B.告警信息分析

     C.断开业务

     D.替换部件

 Page 94



习题答案：

判断题：1.T

单选题：1.BD

Thank you

www.huawei.com

HC120920006 统一存储系统性能与优化




HC120920006
统一存储性能与优化
www.huawei.com
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



目标

- 学完本课程后，您将能够:
  - 熟悉系统性能指标
  - 熟悉影响性能的关键因素及技术
  - 熟悉性能诊断和调优方法
  - 熟悉性能测试工具和方法
  - 熟悉SAN存储系统常见性能故障排除

 Page 2

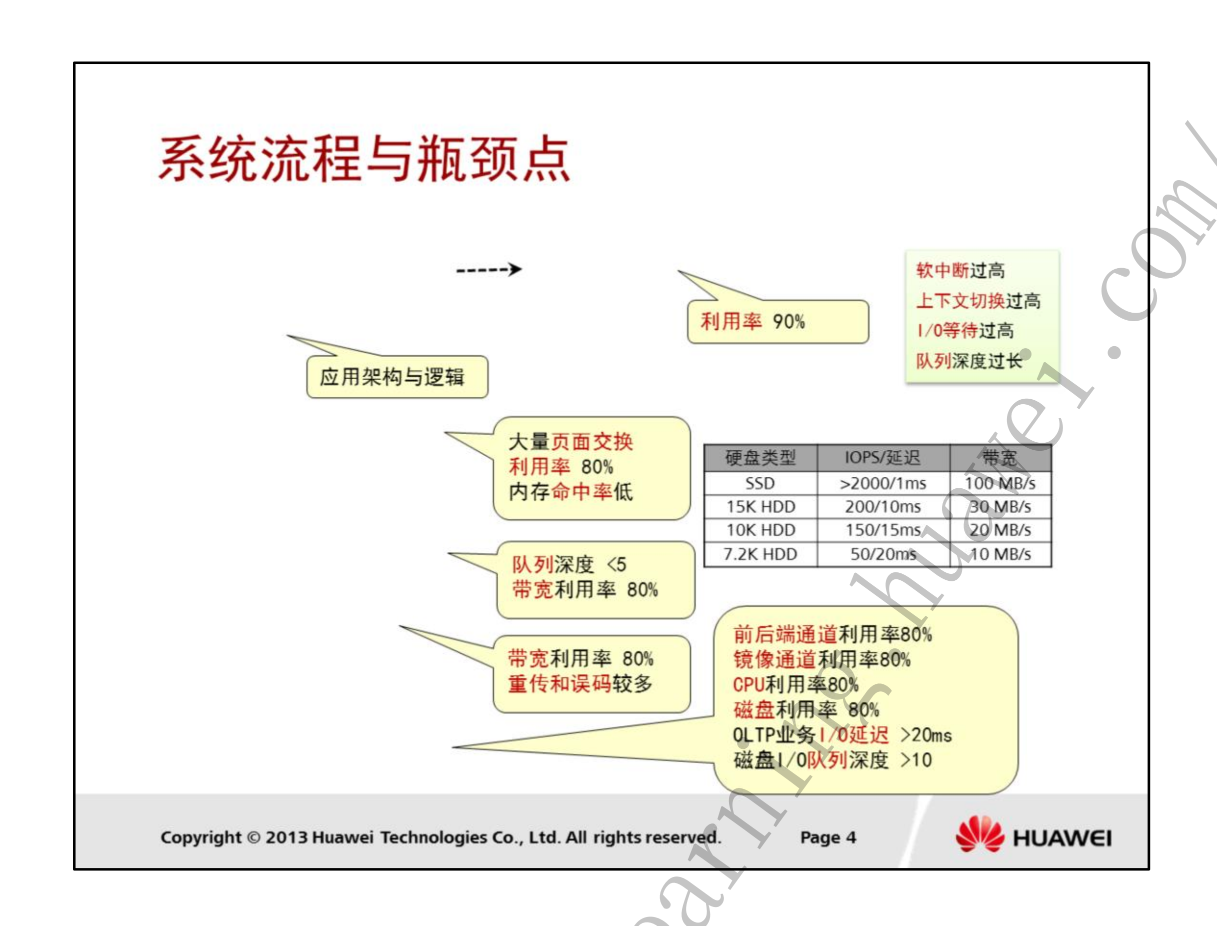




目录

1. **性能调优概述**
2. 性能指标
3. 影响性能的关键因素及技术
4. 性能诊断和调优
5. 性能测试工具和方法
6. SAN存储系统常见性能故障排除





如左图所示，逻辑模块为业务系统中的软件部分，而硬件资源为业务系统中的硬件部分。当用户请求下发时，业务系统按至上而下的顺序由各逻辑模块对IO请求进行响应和处理。

硬件资源为各逻辑模块的正常工作提供支持。一旦逻辑模块在设计和配置上不合理，会直接引起硬件资源衰竭、响应缓慢或受到限制，导致性能问题的出现。

在性能调优时，我们需要在明确性能需求的前提下，以系统IO流程为线索，确定具体哪种硬件资源成为瓶颈，是由什么原因导致的，再结合逻辑模块的设计与配置进行优化。

CPU作为整个系统的大脑，是最重要的资源，它被如下应用所占用：

- OS内核运行及系统调用用户应用程序运行系统触发的软、硬中断处理
- 系统上下文切换
- Iowait
- 虚拟机运行

由于CPU资源工作繁忙，很容易成为整个系统的主要瓶颈。CPU利用率是判定CPU是否成为瓶颈的主要依据．如右图所示，当CPU利用率大于90%时，意味着CPU已成为系统瓶颈．需结合CPU性能分析工具对导致CPU出现瓶颈的应用进行详细分析和优化。



- 必备条件

1. 熟悉整个系统的架构与业务特性
2. 知道性能问题发生的现象与触发条件
3. 熟练掌握各模块的性能监控工具



- 性能调优成本分析
  - 一套业务系统的性能是由系统中的性能短板决定的，每套业务系统在不同的业务特性下都存在短板。
  - 任何优化都存在限制，做超出实际需求的优化都是时间和金钱的浪费。


系统调优流程
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 7
HUAWEI



目录



常用术语解释

术语	解释
慢盘	磁盘的IO响应很慢，导致该磁盘的读写性能明显低于其他磁盘
脏数据	在存储的Cache中有，但是在存储的磁盘上没有的数据
写惩罚	为确保数据一致性，为完成某个IO的写操作而导致的多余的读操作
寻道延迟	为完成读写操作，磁头在电路控制下径向移动到指定磁道
旋转延迟	为完成读写操作，磁盘高速旋转，等待盘片旋转到确定的位置所需要花费的时间



性能指标简介

IOPS

- I/O per second
- 每秒钟存储可以处理的IO数目

带宽

- 常以MB/s为单位
- 即每秒存储可以处理的数据量

响应时间

- 从IO下发到IO处理完成的时间
- 常以ms为单位
- 常用指标：平均响应时间、最大响应时间

波动率

- 衡量方式：最大值、最小值、均方差
- 最常用的方式：均方差/平均值 * 100%

 Page 10 HUAWEI

IOPS、带宽和响应时间是IO性能最核心的性能指标，直接反映当前业务的IO性能状况

波动率是衍生性能指标，对于某些需要恒定性能输出的业务，性能波动过大会直接导致业务故障

- IOPS与带宽的关系

  带宽 = IOPS * 平均IO大小

- IOPS测试常用小IO，带宽测试常用大IO
- IOPS与响应时间的关系

  number of worker ×number of outstanding IO ×1000 / IOPS = Average I/O Response Time (ms)

- IOPS=并发系数/IO平均响应时间

什么是性能问题？

1. 性能数据波动大，稳定性不够
2. 性能值在不同版本之间变化明显
3. 性能数据不能满足测试的需求
4. IO延迟时间长，用户明显感觉业务响应慢

存储性能调优思路

- Step1 确保测试环境正确、稳定
- Step2 判断IO已经正常到达存储前端，性能瓶颈在存储端
- Step3 确保存储配置在对应业务类型下是性能最优的
- Step4 在存储端通过命令和工具进行定位和调优





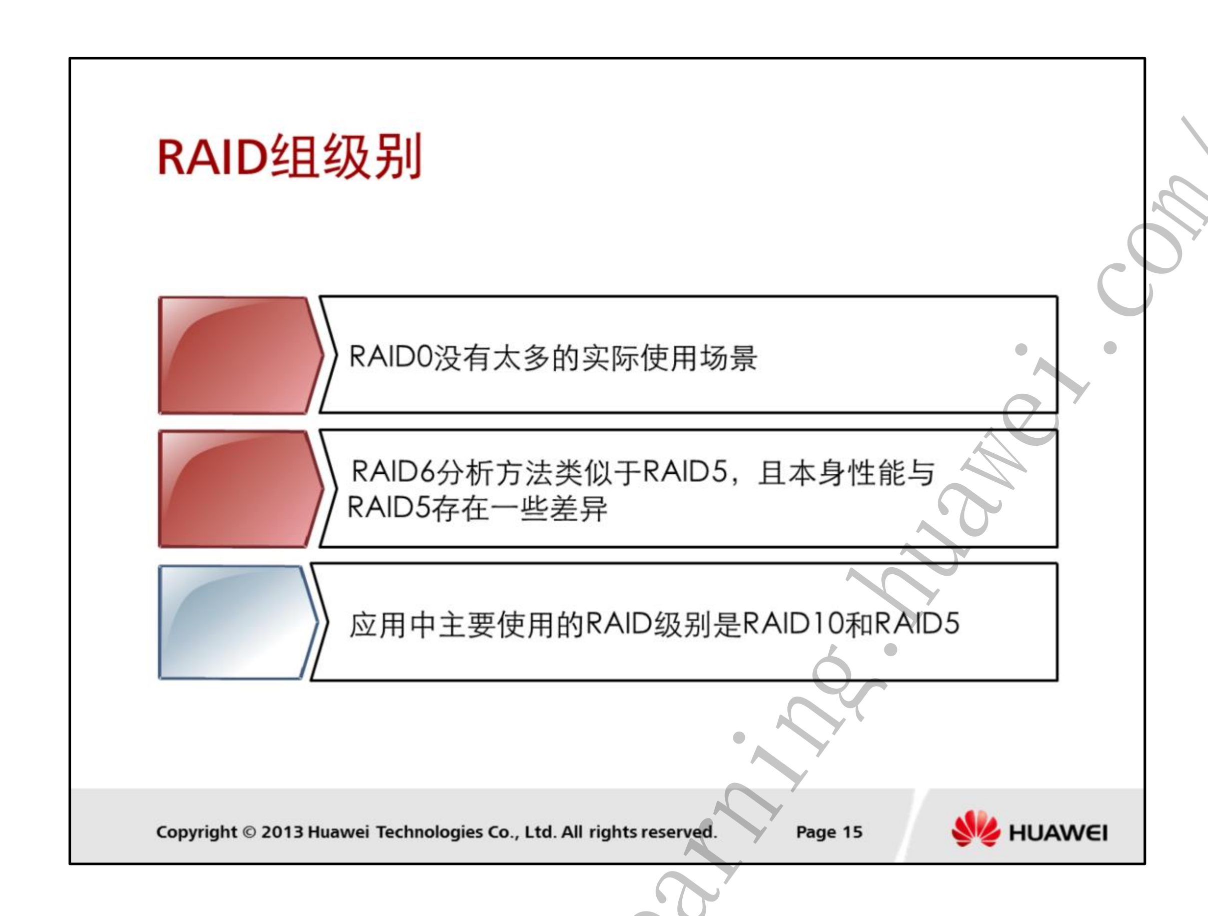
目录



存储配置对性能的影响

- RAID组级别
- 成员盘个数
- 分条深度
- 读写策略
- 高低水位
- LUN归属





RAID组级别作为一种算法，有机的将各个分散的磁盘通过某种方式组合到了一起，并进行条带化的处理。使得多个磁盘可以同时有效的工作，有效的提升系统整体处理IO的能力，提高了数据的安全性；

RAID组级别— RAID5

- 满分条写

需要修改整个分条的所有分条单元，校验数据由新写入的数据计算

- 读改写（小写）

要写入的分条单元数目不足磁盘数目的一半，新的校验数据=老数据XOR 新数据XOR老校验数据

- 重构写（大写）

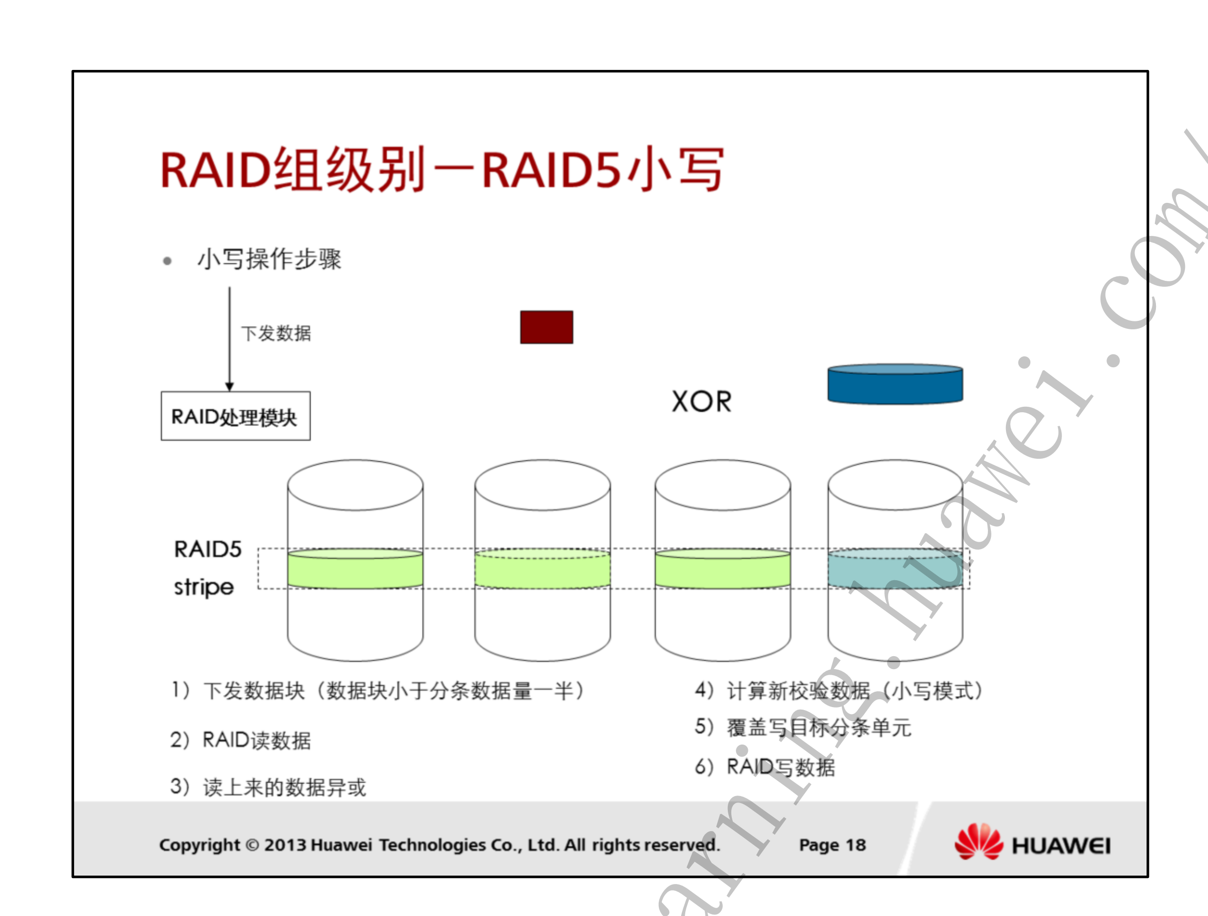
要写入的分条单元数目超过磁盘数目的一半，新的校验数据=新的数据XOR不需要修改的分条

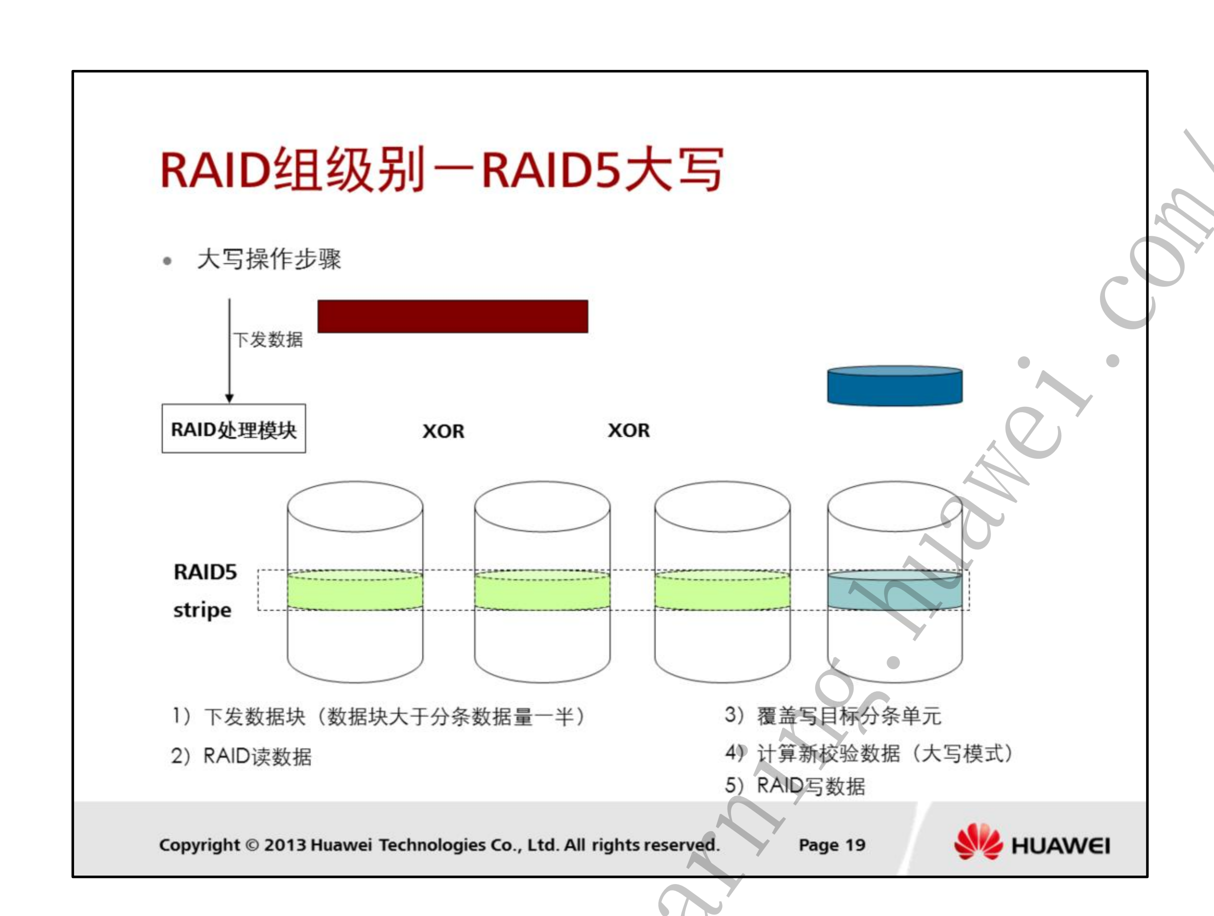


小写过程是：依次读取需要修改的分条上的旧数据和旧的校验数据，根据如下公式计算新的校验数据，并将新数据和新的校验数据写盘。

RAID组级别—RAID5

- 最小操作单元：下发IO的大小；
- 小写优化：


RAID组级别－RAID5小写
小写操作步骤
下发数据
RAID处理模块
XOR
RAID5
stripe
1）下发数据块（数据块小于分条数据量一半）
2）RAID读数据
3）读上来的数据异或
4）计算新校验数据（小写模式）
5）覆盖写目标分条单元
6）RAID写数据
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 18
HUAWEI


RAID组级别—RAID5大写
大写操作步骤
下发数据
RAID处理模块
XOR
XOR
RAID5
stripe
1）下发数据块（数据块大于分条数据量一半）
2）RAID读数据
3）覆盖写目标分条单元
4）计算新校验数据（大写模式）
5）RAID写数据
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 19
HUAWEI

RAID组级别—RAID10

因为不需要生成校验位信息，因此RAID10下的最小操作单元：IO数据块的大小

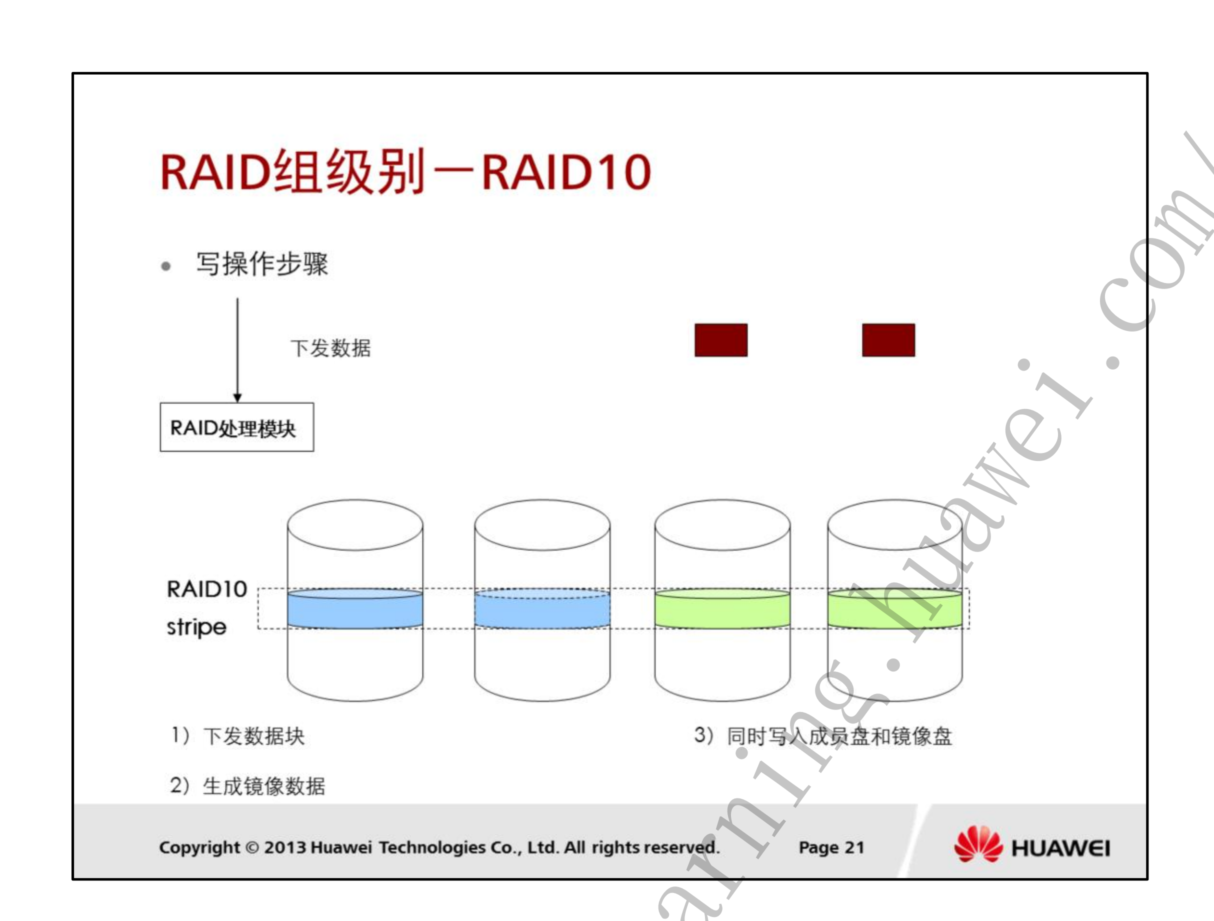
由于每个镜像盘组内的数据完全相同， 因此RAID10下的读操作从这其中的N/2块磁盘即可获取数据

对写操作，需要对镜像组内的所有盘都写该数据，分别写到镜像组内的两块磁盘上

Page 20



前提条件：假设有N块磁盘组成了一个RAID10（N为偶数），且该RAID10的镜像盘个数为2。


RAID组级别－RAID10
写操作步骤
下发数据
RAID处理模块
RAID10
stripe
1）下发数据块
2）生成镜像数据
3）同时写入成员盘和镜像盘
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 21
HUAWEI

RAID组级别—顺序读性能对比

- RAID5：除去校验盘以后，对一个分条而言，有（N-1）块磁盘可以同时提供数据；
- RAID10：有N块盘可以同时提供数据，但由于镜像组内的数据是完全相同的，因此对同一个分条来说，其实上只有N/2块盘可同时提供不同的数据；
- 由于同一个分条参与读操作的硬盘RAID5多于RAID10，因此RAID5读性能高。



RAID组级别－顺序性能对比

RAID5与RAID10顺序性能对比

- 持续顺序读性能
  - RAID5性能高于RAID10
- 持续顺序写性能
  - 前端压力足够的情况下，RAID5性能明显高于RAID10

顺序读带宽

顺序写带宽

 Page 23 HUAWEI

- 持续顺序读性能
  - RAID5：除去校验盘以后，对一个分条而言，有（N-1）块磁盘可以同时提供数据；
  - RAID10：有N块盘可以同时提供数据，但由于镜像组内的数据是完全相同的，因此对同一个分条来说，其实上只有N/2块盘可同时提供不同的数据；
  - 由于同一个分条参与读操作的硬盘RAID5多于RAID10，因此RAID5读性能略高
- 持续顺序写性能
  - RAID5：除去校验盘以后，对一个分条而言，有（N-1）块磁盘可以同时接收写下来的数据；同时需要计算以生成该分条新的校验信息；
  - RAID10：假设镜像盘的个数为2，那么当IO下发的时候，对于同一个镜像组内的磁盘，实际上是写一份数据再写一份镜像数据，两份数据完全一样，即只有N/2块磁盘在接收不同的IO；
  - 前端压力较小的时候：由于RAID5会耗费时间计算校验信息，因此RAID5的持续顺序写带宽会小幅优于RAID10；
  - 前端压力较大的时候：此时后端的环路带宽是性能的瓶颈点。此时，RAID5的写带宽会优于RAID10的写带宽，且RAID10的写带宽会略高于RAID5的写带宽值的一半。

RAID组级别—随机性能对比

RAID5与RAID10随机性能对比

- 持续随机读性能
  - RAID5与RAID10性能基本相当
- 持续随机写性能
  - RAID5性能明显低于RAID10

持续随机读IOPS

持续随机读IOPS

IOPS

2000.00 1800.00 1600.00 1400.00 1200.00 1000.00 800.00 600.00 400.00 200.00 0.00

Raid5 Raid10

Raid组级别

持续随机写IOPS

持续随机写IOPS

IOPS

1000.00 900.00 800.00 700.00 600.00 500.00 400.00 300.00 200.00 100.00 0.00

Raid5 Raid10

Raid组级别

 Page 24 HUAWEI

- 持续随机读性能
  - RAID5和RAID10处理读IO请求的最小操作单元都是IO的大小。
  - 随机业务的IO比较离散，虽然RAID5和RAID10方式能同时提供数据的盘数不同，但性能上并不会有太大的差异。
  - RAID10的随机读性能与RAID5的随机读性能基本相当。
- 持续随机写性能
  - 在通常的情况下，随机写时的IO大小一般不会太大，该大小一般是小于分条单元的大小的。
  - RAID5算法中的“写惩罚”会导致多余的读操作；RAID10不存在这个问题。
  - 结论：RAID10的随机写性能会明显的好于RAID5的随机写性能。

RAID组级别—顺序写性能对比

- RAID5：除去校验盘以后，对一个分条而言，有（N-1）块磁盘可以同时接收写下来的数据；同时需要计算以生成该分条新的校验信息；
- RAID10：假设镜像盘的个数为2，那么当IO下发的时候，对于同一个镜像组内的磁盘，实际上是写一份数据再写一份镜像数据，两份数据完全一样，即只有N/2块磁盘在接收不同的IO；
- 前端压力较小的时候：由于RAID5会耗费时间计算校验信息，因此RAID5的持续顺序写带宽会小幅优于RAID10，前端压力较大的时候：此时后端的环路带宽是性能的瓶颈点。此时，RAID5的写带宽会优于RAID10的写带宽，且RAID10的写带宽会略高于RAID5的写带宽值的一半。

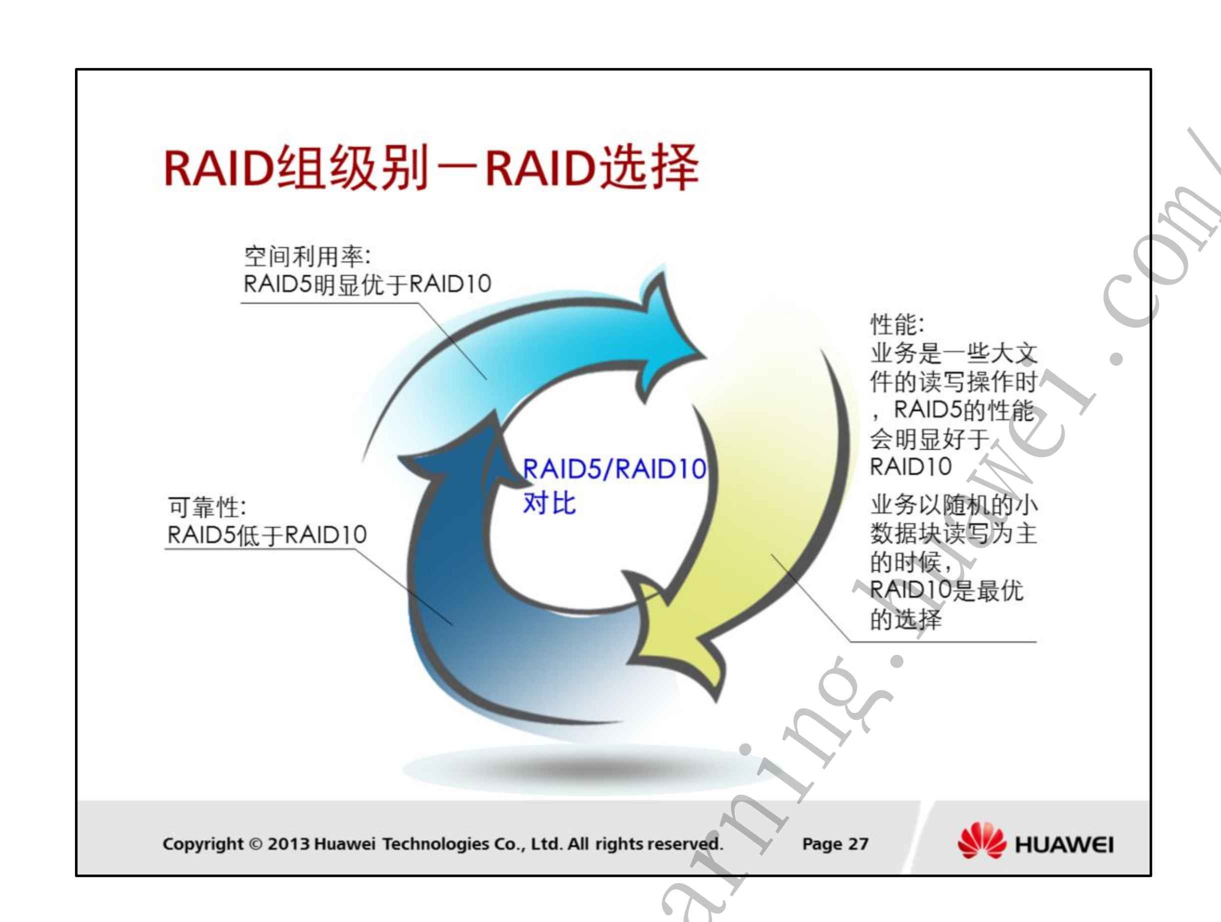


RAID组级别—随机性能对比

- 持续随机读性能
  - RAID5和RAID10处理读IO请求的最小操作单元都是IO的大小。
  - 随机业务的IO比较离散，虽然RAID5和RAID10方式能同时提供数据的盘数不同，但性能上并不会有太大的差异。
  - RAID10的随机读性能与RAID5的随机读性能基本相当。
- 持续随机写性能
  - 在通常的情况下，随机写时的IO大小一般不会太大，该大小一般是小于分条单元的大小的。
  - RAID5算法中的“写惩罚”会导致多余的读操作；RAID10不存在这个问题。
  - RAID10的随机写性能明显的好于RAID5的随机写性能。

 Page 26




RAID组级别－RAID选择
空间利用率:
RAID5明显优于RAID10
性能:
业务是一些大文件的读写操作时，RAID5的性能会明显好于RAID10
业务以随机的小数据块读写为主的时候，RAID10是最优的选择
RAID5/RAID10
对比
可靠性:
RAID5低于RAID10
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 27
HUAWEI

存储配置对性能的影响

- RAID组级别
- **成员盘个数**
- 分条深度
- 读写策略
- 高低水位
- LUN归属



随机业务与成员盘个数的关系

- 一块磁盘只能承担一定数目的随机IO访问量，这是由磁盘的平均旋转延迟和寻道延迟所决定的。
- 对于随机读写业务的测试，假设存储设备前端IO的压力足够，当增加RAID组内的成员盘个数的时候，由于有更多的磁盘来分担下发的随机读写请求，随机读写IOPS性能数值也是逐渐增加的。



控制器内存CHUNK介绍

- CHUNK是一种数据结构。在控制器Cache模块，系统会将满足一定规则的IO装载在一个CHUNK中，每次将按照一个CHUNK的大小为单位将IO数据下发到RAID模块，并最终刷盘。
- CHUNK就是一个量筒。对于某一个确定的量筒，其最大的容量是确定不变的。如果需要量取超过该量筒容量限制的液体，那么就肯定需要操作多次；而每次也可以量取不超过其容量限制的液体。CHUNK也是这样的，CHUNK的大小就是对应于某个量筒的最大量取容量，不同配置下的CHUNK大小可能不同，但是最大不超过1MB。



控制器内存CHUNK结构

1chunk≤16chunkunit

chunkunit
chunkunit
⋮
chunkunit
chunkunit
chunkunit

1chunkunit=16page

page
page
⋮
page
page
page

1page=8sector

sector
sector
⋮
sector
sector
sector

- CHUNK的容量上限是：512B×8×16×16＝1MB
- CHUNK-UNIT的大小是：512B×8×16＝64KB

 Page 31 HUAWEI

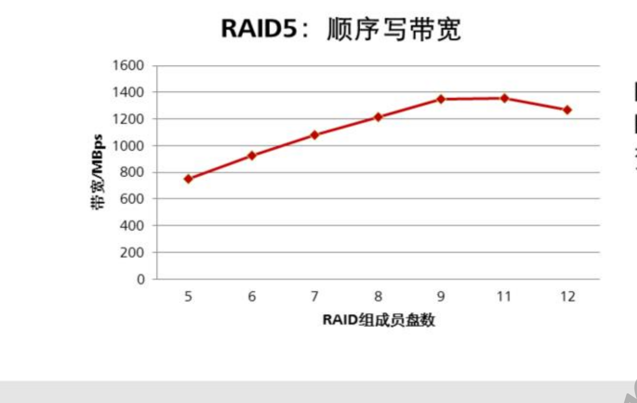
控制器内存CHUNK大小的确定

- CHUNK的大小是与分条的大小密切相关的。
- 分条的大小 = 分条深度的大小 × 有效盘数。
- 若分条的大小大于1MB时，那么CHUNK的大小就是1MB。
- 当分条的大小小于1MB时，CHUNK的大小会首先按照分条的大小对齐，然后再按照CHUNK-UNIT（64K）的大小对齐，最终得到CHUNK的大小。



RAID5：成员盘个数与持续顺序写带宽

- 顺序写带宽的主要影响因素：CHUNK的大小和分条的大小。
- 举例：128KB分条深度（固定不变）



RAID5持续顺序写带宽随RAID组成员盘个数增加的变化趋势图



CHUNK大小越大，对持续顺序写操作来说，意味者每次下发的数据量越多，此时只要满足CHUNK对齐分条的原则，则CHUNK大小越大，写性能就越好。

成员盘个数增加，CHUNK和分条的大小都不断增大，但始终保持两者相等，直至成员盘个数为9盘时，此时分条CHUNK的大小均为128KB×（9-1）＝1MB。

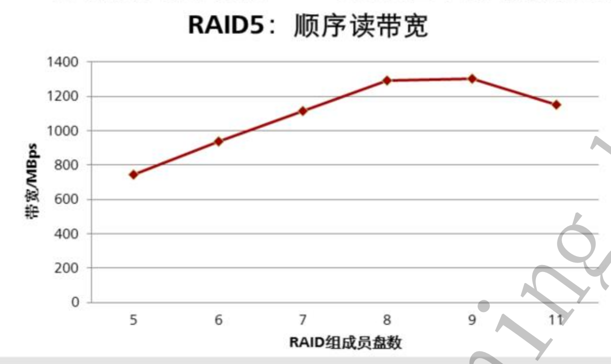
继续增加成员盘个数到11盘，此时分条的大小为128KB×（11-1）＝1280KB，但由于分条的大小已超过CHUNK的最大大小值，则此时CHUNK大小为1MB。

在这种情况下，对于持续顺序写操作，会出现一个分条请求被拆分为两个CHUNK来完成的情况，因此此时的写性能出现了较为明显的下降。

RAID5：成员盘个数与持续顺序读带宽

- 持续顺序读带宽的主要影响因素：分条的大小和预取算法的不同所引起的预取值不同。
- 举例：128KB分条深度、智能预取。

RAID5持续顺序读带宽随RAID组成员盘个数增加的变化趋势图

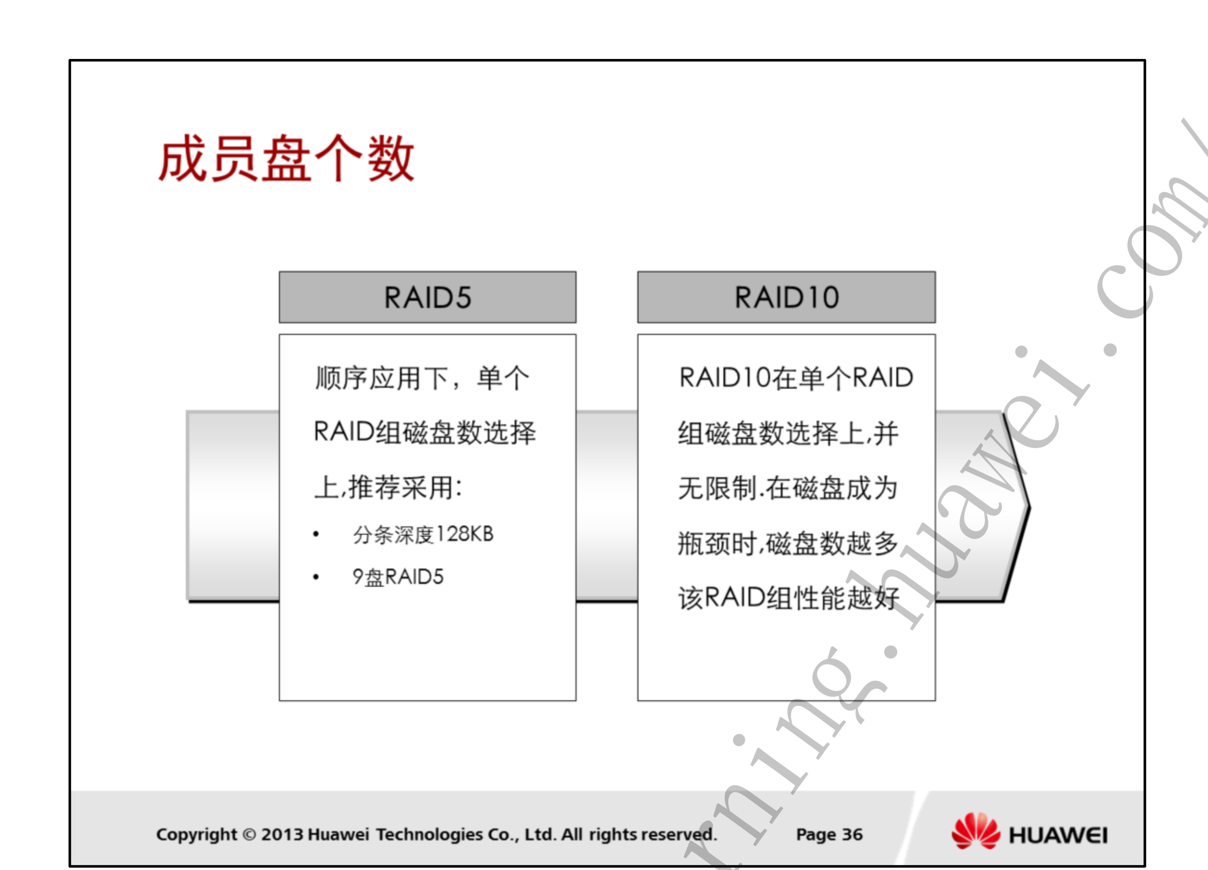


RAID10：成员盘个数与持续顺序读写业务

- RAID10并没有校验信息，不管是何种业务，一个RAID10下的镜像组之间可以完全认为是独立的。
- 因此一个RAID10下更多的磁盘数目实际上可以分担更多的IO，即可提升持续顺序读写带宽的性能。

 Page 35




成员盘个数
RAID5
顺序应用下，单个RAID组磁盘数选择上,推荐采用:
分条深度128KB
9盘RAID5
RAID10
RAID10在单个RAID组磁盘数选择上,并无限制.在磁盘成为瓶颈时,磁盘数越多该RAID组性能越好
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 36
HUAWEI

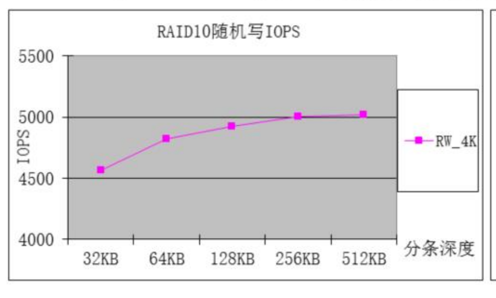
存储配置对性能的影响

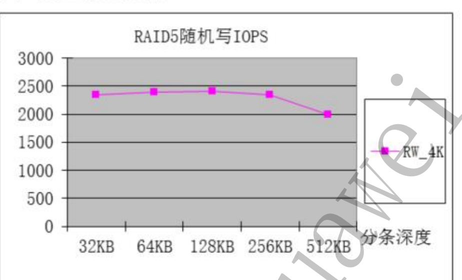
- RAID组级别
- 成员盘个数
- **分条深度**
- 读写策略
- 高低水位
- LUN归属



分条深度—随机写性能

RAID5和RAID10下分条深度变化随机写性能规律



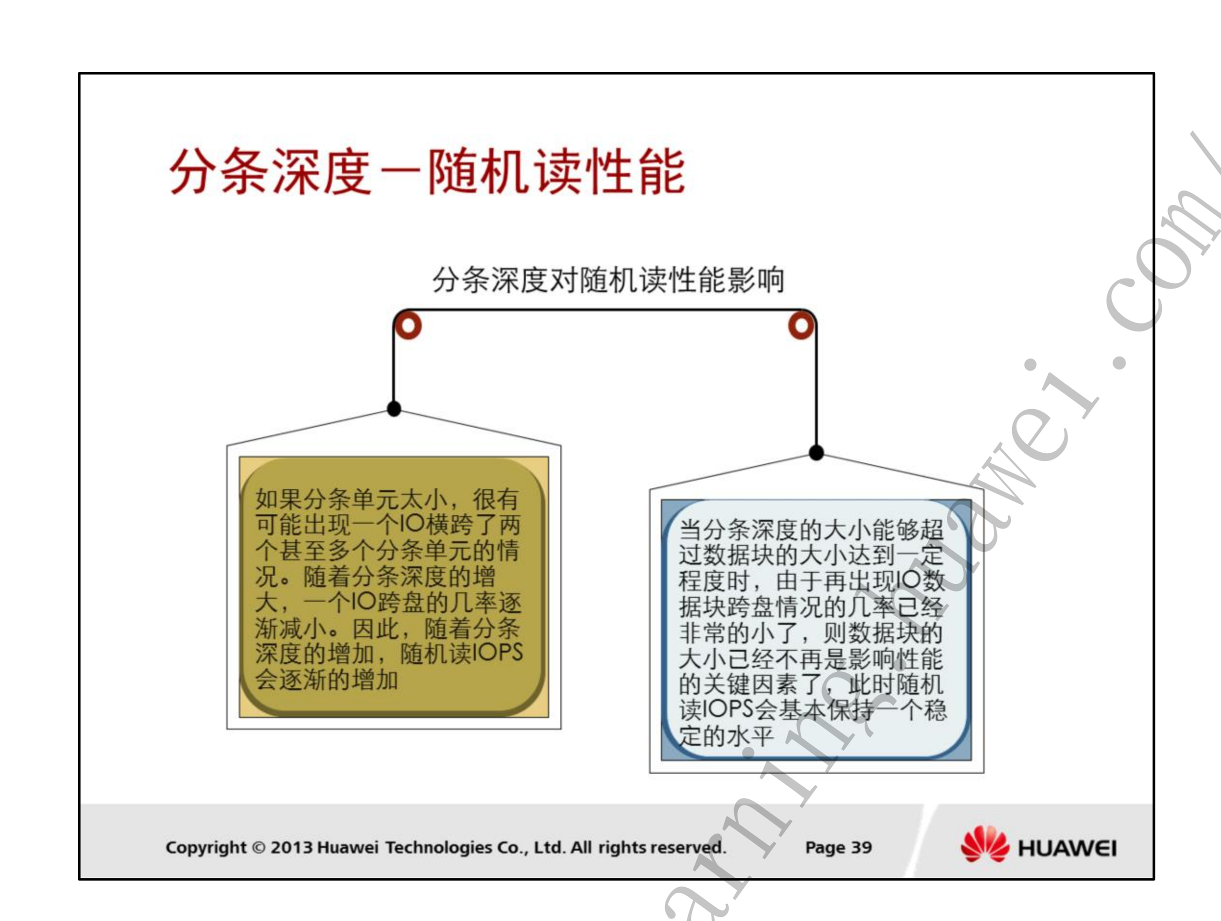


RAID5规律：随着分条深度的增加，随机写IOPS先会不断的增加，到达一定程度之后，随机写IOPS会不断的递减;

RAID10规律：随着分条深度的增加，随机写IOPS不断的增长，当分条深度增大到一定程度后，随机写IOPS保持一个较为稳定的状态;

 Page 38 

- RAID5随机写性能与分条深度变化规律分析
  - 分条深度的大小从两个方面影响RAID5下的随机写IOPS：IO的分布以及写惩罚；
  - 当分条深度较小时，分条深度主要影响IO的分布，这时出现一个IO横跨两个或多个分条单元的几率比较大；为了写一个IO，需要操作多个磁盘，因此随机写IOPS的性能会比较低；随着分条深度的增加，一个IO落到单个分条单元中的几率逐渐的增大，因此随机写IOPS值也就随之得到了提升；
  - 随着分条深度的进一步增加，一个IO出现横跨分条的几率更加小了；此时，写惩罚成了影响性能的关键因素，因此此时随分条深度的增加，随机IOPS的性能反而下降了；
- RAID10随机写性能与分条深度变化规律分析
  - 分条深度的大小从IO的分布影响RAID10下的随机写IOPS；
  - 随分条深度的增大，一个IO跨盘的几率逐渐的减小，因此随机写IOPS呈现逐渐上升的趋势；
  - 当分条深度增加到一定程度之后，再继续增加分条深度的大小并不能有效的增加一个IO落入同一个分条单元的几率，因此随机写IOPS的增长趋势逐渐变缓；


分条深度－随机读性能
分条深度对随机读性能影响
如果分条单元太小，很有可能出现一个IO横跨了两个甚至多个分条单元的情况。随着分条深度的增大，一个IO跨盘的几率逐渐减小。因此，随着分条深度的增加，随机读IOPS会逐渐的增加
当分条深度的大小能够超过数据块的大小达到一定程度时，由于再出现IO数据块跨盘情况的几率已经非常的小了，则数据块的大小已经不再是影响性能的关键因素了，此时随机读IOPS会基本保持一个稳定的水平
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 39
HUAWEI

存储配置对性能的影响

- RAID组级别
- 成员盘个数
- 分条深度
- **读写策略**
- 高低水位
- LUN归属



读策略

- 读预取的作用
  - 在处理一个读IO请求时，从磁盘侧按顺序读取除该IO数据以外更多的数据，预先缓存到Cache中，以便下一个顺序读IO请求到达时，可直接在Cache中获取，得到更高的性能表现。
  - 当读IO很随机时，不当的读预取策略会给存储系统带来额外的资源开销，不但无法保证后续IO在Cache中的命中，而且会带来性能的降低。
- 四种预取算法：
  - 固定预取、可变预取、智能预取和不预取。



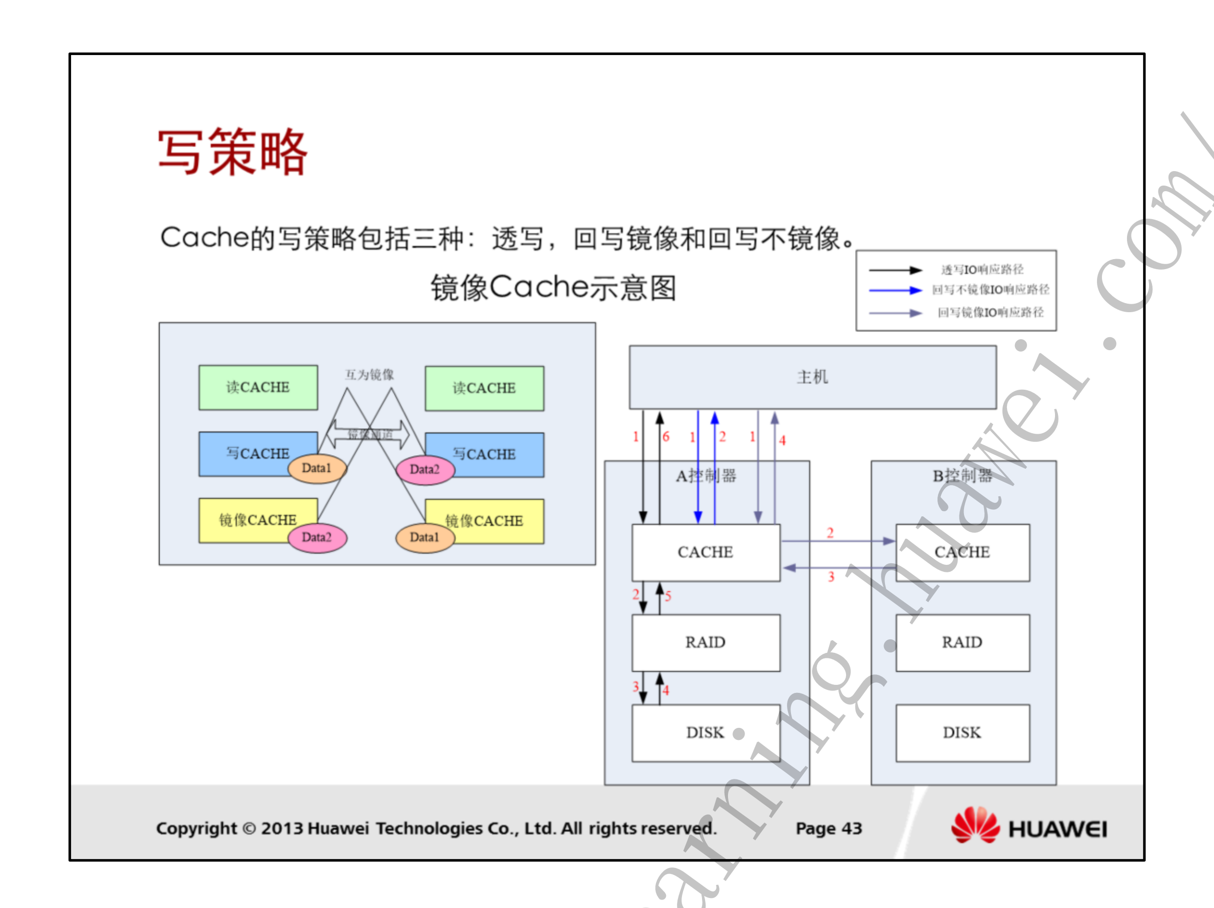
读策略

预取策略	策略说明	适用场景
固定预取	按系统设置的固定预取窗口大小进行预取	1. 适用于业务单一，IO顺序性强的场景，如：DSS和数据备份等业务 2. 建议预取窗口大小与LUN分条对齐
可变预取	按系统设置的倍数，预取：读IO长度*倍数	适用于IO顺序性强，但IO大小呈规律性变化的业务
智能预取	按IO顺序程度，动态起停预取，动态调整预取窗口大小	适用于IO特性比较复杂的场景，如：OLTP应用中数据文件访问
不预取	减少因预取无效数据引起系统额外开销	适用于可明确的IO随机场景

 Page 42



- 固定预取:当某个IO不命中的时候，需要下到磁盘取数据，并顺便取与该IO地址连续的一定量的IO上来，该数量是事先设定的，其取值范围是0到1024KB，当选择的预取数值为0时实际上就是不预取；
- 可变预取:可变预取的预取值可以在0到65535中进行选择；当选择为0的时候，其实就是不预取；
- 智能预取:智能预取算法可动态分析当前业务IO顺序性情况和IO大小规律，判定是否起停预取和预取窗口大小；



- 透写与回写
  - 透写：每一个写IO请求都直接下到磁盘，整体性能的表现受限于磁盘本身所能提供的性能，主要取决于磁盘的转速，寻道时间等关键的参数。
  - 回写：每一个写IO到达Cache就代表写入成功；然后通过Cache层面对数据进行充分的整合与调度，再统一下发到磁盘。由于写入Cache的速度会快很多，因此可以得到比透写时更好的性能表现。
- 镜像与不镜像
  - 镜像与不镜像的差别在于：当一个写IO从A控制器下发，当到达A控制器的Cache之后，是否需要将该IO再写入到B控制器的Cache作备份；
  - 如果采用了回写镜像的方式，由于进行了额外的两个控制器之间的交互以及数据的传输，因此总体的性能表现不会比回写不镜像的方式好。

写策略

- 需仔细评估实际业务对性能和可靠性的需求，选择恰当的写策略：

写策略	可靠性	性能
透写	高	低
回写镜像	中	中
回写不镜像	低	高

 Page 44

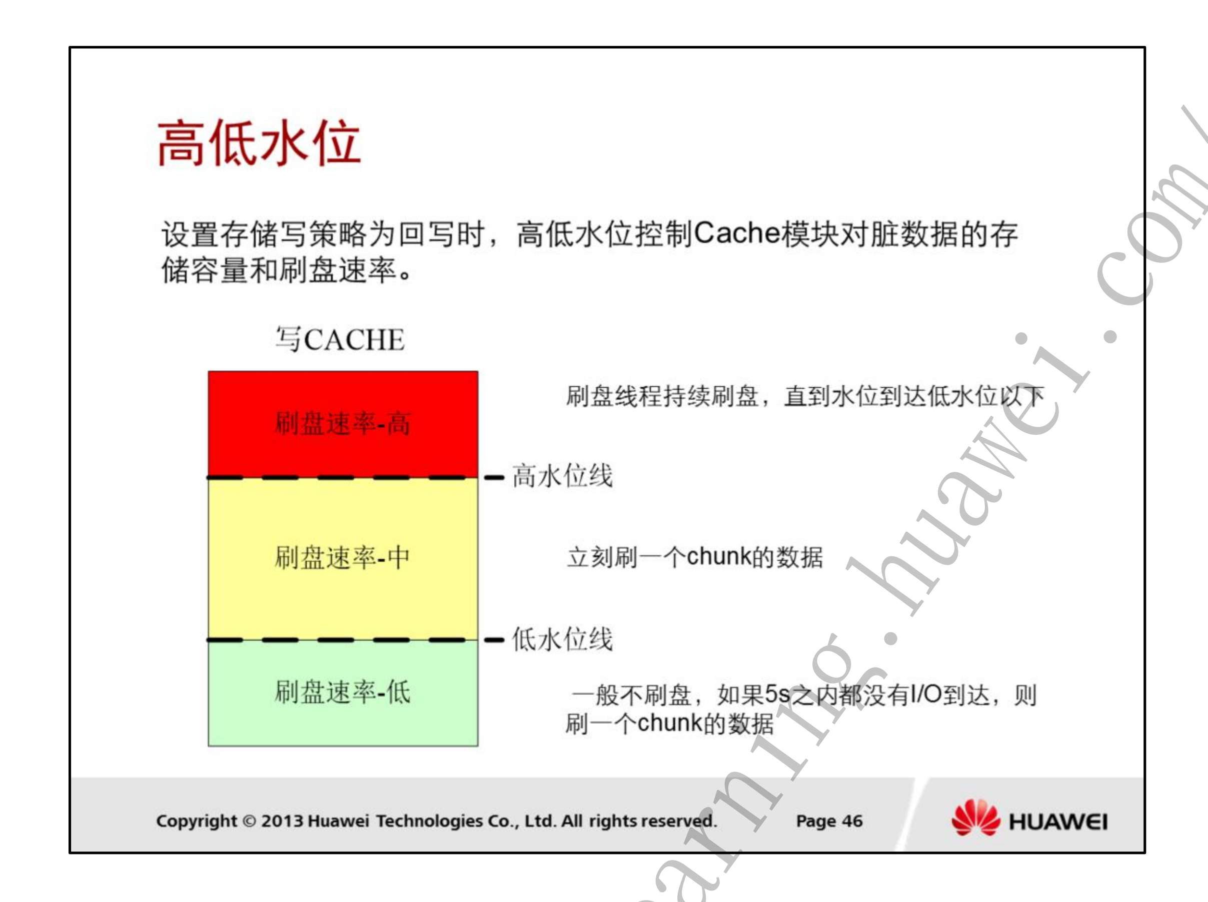


对于随机的写业务来说：由于业务且流量较小，性能的瓶颈主要在系统资源的分配、后端磁盘的个数、磁盘的转速等因素上，镜像并不是写性能的关键的因素。不管是否采用了镜像的方式，随机写性能几乎没有区别。如果出现镜像时的随机写性能略微优于不镜像时的随机写性能也是正常的现象。

存储配置对性能的影响

- RAID组级别
- 成员盘个数
- 分条深度
- 读写策略
- **高低水位**
- LUN归属





- 将脏数据缓存在Cache中可使IO得到充分整合与调度，降低延迟，提升性能；但当写Cache中缓存的IO脏数据的总量达到一定上限时，就需要加快数据刷盘的速度，避免由于写Cache缓存了过多脏数据不能接收前端下发的新的写IO请求；
- 默认情况下，系统设置的低水位值为20，在运行实际的业务，特别是随机写业务的时候，可适当的提升低水位值，如设置为40或者50；
- 高水位需根据实际业务的下发情况进行调整，重点关注调整后的性能变化状况和Cache命中率状况，针对随机写业务建议设置在80左右；
- 刷盘线程每隔1s启动一次

高低水位特性

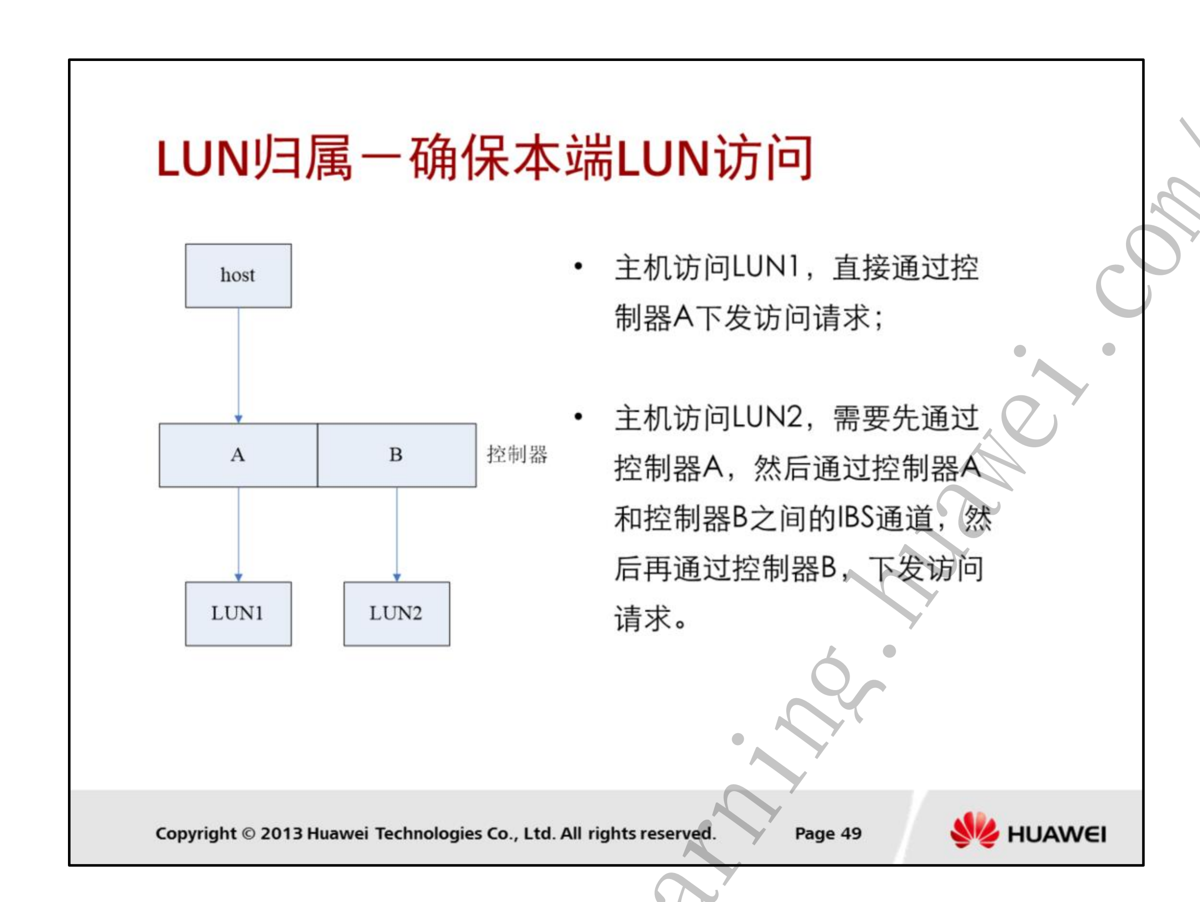
- Cache中缓存的数据与低水位值所对应的容量持平或者低于该容量的时候，这些数据被刷盘的几率相对是较小的，即这些数据保存在Cache中的几率相对较大。
- IO停留在Cache中的时间与低水位值的设置密切相关。当将低水位值设置得高一些，就可以使得下发的IO获得更多的整合的机会，从而提升随机写的性能。
- 默认情况下，系统设置的低水位值为20，在运行实际的业务，特别是多路顺序小IO和SPC-1模型的OTLP业务时，可适当的提升低水位值，如设置为40或者50。



存储配置对性能的影响

- RAID组级别
- 成员盘个数
- 分条深度
- 读写策略
- 高低水位
- **LUN归属**





- 对LUN1的访问即是本端访问的方式。
- 对LUN2的访问即是对端访问的方式。
- 对端访问的方式必然要通过板间的IBS通道，因此必然受到IBS通道的限制，所以会出现其读写的性能受到影响的情况。因此，我们在归属LUN的时候，需要注意将该LUN归属给将访问其的主机所连接的控制器上。



目录



Linux常用性能监控命令

```
[root@localhost ~]# iostat -kx 1
Linux 2.6.35.6 (localhost.localdomain)  01/10/2012      _x86_64_        (24 CPU)

avg-cpu:  %user   %nice %system %iowait  %steal   %idle
           0.01    0.00    0.10    0.02    0.00   99.88

Device:         rrqm/s   wrqm/s     r/s     w/s    rkB/s    wkB/s avgrq-sz avgqu-sz   await  svctm  %util
sda               0.29     0.27    0.39    0.16    12.48     1.62    51.58     0.01   11.15   1.15   0.06
dm-0              0.00     0.00    0.59    0.40    12.17     1.62    27.60     0.01   12.62   0.62   0.06
dm-1              0.00     0.00    0.02    0.00     0.07     0.00     8.00     0.00    1.70   0.94   0.00
dm-2              0.00     0.00    0.02    0.00     0.06     0.00     7.93     0.00    0.75   0.68   0.00
sdb              22.62    50.47    8.42   18.01   251.32   561.73    61.53     0.09    3.58   1.66   4.38
sdc              14.69     0.03    0.61    0.73    61.56   182.83   364.62     0.00    2.03   1.97   0.26
sdd               0.36   842.36    0.20   52.55     4.07  7068.77   268.14     0.12    2.25   0.66   3.47
sde               0.00     0.00    0.11    0.07     4.34     5.07   102.65     0.00    0.46   0.46   0.01
sdg               0.00     0.00    0.11    0.07     4.35     5.07   104.55     0.00    0.48   0.48   0.01
sdf               0.36     0.50    0.20    0.13     4.07     4.94    54.72     0.00    0.40   0.28   0.01
dm-3              0.00     0.00    0.63  923.79     4.66  7290.72    15.78     5.15    5.57   0.06   5.40
sdh               0.00     0.00    0.00    0.00     0.00     0.00     8.00     0.00    1.22   1.22   0.00
dm-4              0.00     0.00    0.63    0.66     4.65     5.24    15.33     0.00    0.77   0.10   0.01
```

- 顺序业务
  - util%应接近100%
  - rkb/s、wkb/s应达到通道理论带宽
  - avgrq-sz应等于上层业务的Block大小
- 随机业务
  - r/s、w/s应等于理论计算所得到的IOPS值
  - avgqu-sz应达到适当的值
  - await应小于30ms。

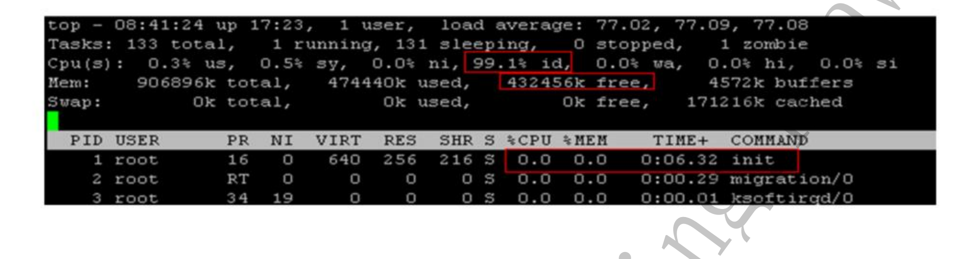
 Page 51



- rrqm/s：平均每秒钟读IO合并的次数
- wrqm/s：平均每秒钟写IO合并的次数
- r/s：平均每秒钟读IO操作次数
- w/s：平均每秒钟写IO操作次数
- rkB/s：平均每秒钟的读带宽
- wkB/s：平均每秒钟的写带宽
- avgrq-sz：平均IO大小
- avgqu-sz：平均IO队列长度
- await：平均每次I/O操作的总等待时间
- svctm：平均每次I/O操作的存储访问时间
- %util：统计间隔中有多少时间I/O队列是非空的

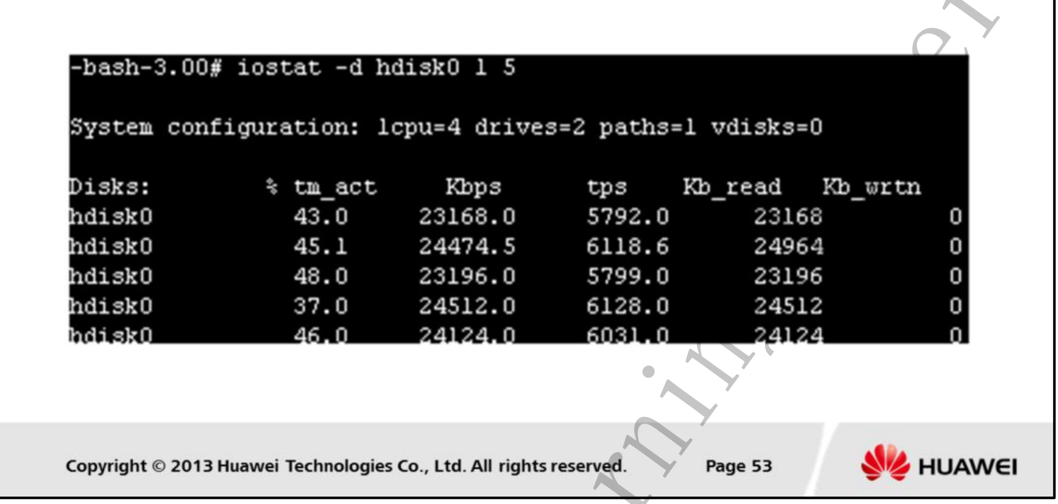
top详细用法

- 第一行中load average显示了过去的1、5、15分钟内运行队列中的平均进程数量
- 第二行显示各种状态的进程的数目
- 第三行显示的是目前CPU的使用情况。
- 第四行显示物理内存的使用情况，
- 第五行显示交换分区使用情况。





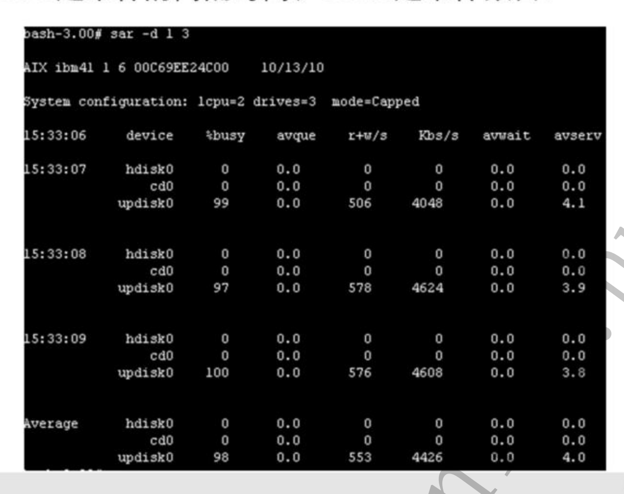
- 第一行中load average显示了过去的1、5、15分钟内运行队列中的平均进程数量
- 第二行显示各种状态的进程的数目
- 第三行显示的是目前CPU的使用情况，包括us用户空间占用CPU百分比、sy内核空间占用CPU百分比、ni用户进程空间内改变过优先级的进程占用CPU百分比(中断处理占用)、id空闲CPU百分比、wa等待输入输出的CPU时间百分比。
- 第四行显示物理内存的使用情况，包括总的可以使用的内存、已用内存、空闲内存、缓冲区占用的内存。
- 第五行显示交换分区使用情况，包括总的交换分区、使用的、空闲的和用于高速缓存的大小。



- %tm_act：表示磁盘利用率，表明了磁盘的繁忙程度。
- Kbps：表示磁盘的数据传输速率。
- tps：表示磁盘的IOPS。
- Kb_read：表示在间隔时间中从磁盘读取的数据量，单位是KB。
- Kb_wrtn：表示在间隔时间中写入磁盘的数据量，单位是KB。
- 虽然该命令没有列出平均IO大小，但是可以通过后面的公式计算得到，平均IO大小=Kbps/tps。

Sar命令

- 通过sar -d [interval] [count]来监控所有设备驱动层disk的性能状况，interval是采样的间隔时间，count是采样数目。



```
bash-3.00# sar -d 1 3

AIX ibm41 1 6 00C69EE24C00    10/13/10

System configuration: lcpu=2 drives=3  mode=Capped

15:33:06     device    %busy    avque    r+w/s    Kbs/s   avwait   avserv

15:33:07      hdisk0      0      0.0        0        0      0.0      0.0
                 cd0      0      0.0        0        0      0.0      0.0
             updisk0     99      0.0      506     4048      0.0      4.1

15:33:08      hdisk0      0      0.0        0        0      0.0      0.0
                 cd0      0      0.0        0        0      0.0      0.0
             updisk0     97      0.0      578     4624      0.0      3.9

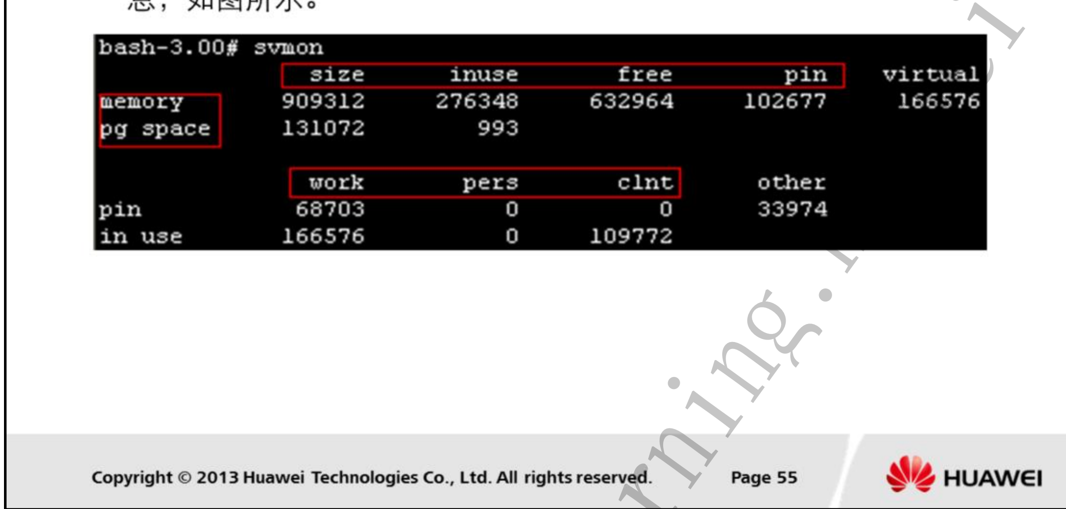
15:33:09      hdisk0      0      0.0        0        0      0.0      0.0
                 cd0      0      0.0        0        0      0.0      0.0
             updisk0    100      0.0      576     4608      0.0      3.8

Average       hdisk0      0      0.0        0        0      0.0      0.0
                 cd0      0      0.0        0        0      0.0      0.0
             updisk0     98      0.0      553     4426      0.0      4.0
```


Page 54



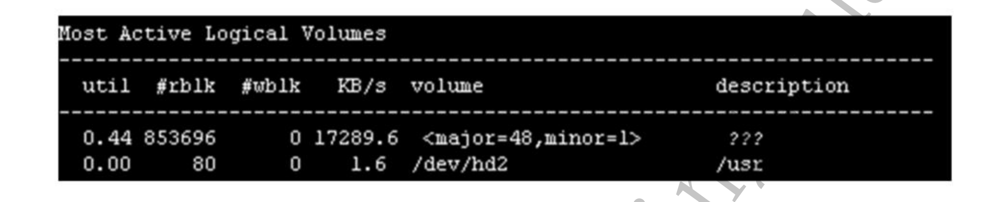
- %busy：表示磁盘利用率，表明了磁盘的繁忙程度。
- avque：表示主机块设备层IO队列深度。
- r+w/s：表示磁盘的IOPS。
- Kbs/s：表示磁盘的数据传输速率。
- avwait：表示主机块设备层的平均IO等待时间。
- avserv：表示主机块设备层的平均IO服务时间。



图中所有数值的单位都是4KB，表示有多少个4K页。其中，size表示总大小，in use表示正在使用中的大小，pin表示正在使用且被锁定的大小，free表示空闲空间的大小。

逻辑卷性能监控

- 可以通过filemon -uo <output_file> -O lv命令来监控逻辑卷的性能，使用trcstop命令来终止filemon，结果保存在output_file中。通过该命令可以了解使用率靠前的几个逻辑卷的性能状况，output_file中会按照使用率从高到低对逻辑卷进行排序，如图所示。



```
Most Active Logical Volumes
------------------------------------------------------------------------
  util  #rblk  #wblk   KB/s  volume                    description
------------------------------------------------------------------------
  0.44 853696      0 17289.6  <major=48,minor=1>         ???
  0.00     80      0     1.6  /dev/hd2                  /usr
```




- util：表示逻辑卷使用率。
- #rblk：表示在间隔时间中从逻辑卷读取的block数，一个block的大小为512B。
- #wblk：表示在间隔时间中写入逻辑卷的block数。
- KB/s：表示逻辑卷的传输速率。
- volume：表示逻辑卷的名称或者设备号。
- description：表示逻辑卷对应的文件系统安装点或者逻辑卷类型。

文件系统性能监控

- 可以通过filemon -uo <output_file> -O lf来监控文件系统的性能，使用trcstop命令来终止filemon，结果保存在output_file中。通过该命令可以了解使用率靠前的几个文件的性能状况，output_file中会按照使用率从高到低对文件进行排序。

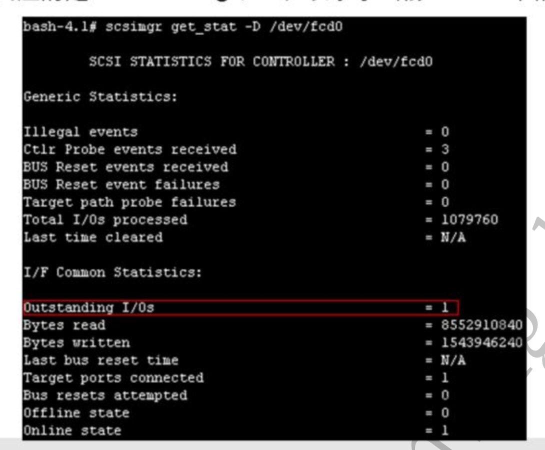
```
Most Active Files
------------------------------------------------------------------------
  #MBs  #opns   #rds   #wrs  file                        volume:inode
------------------------------------------------------------------------
 968.0      1   1936      0  test1                       /dev/lv00:36
 963.0      1      0   1926  null
 157.1      0  40207      0  pid=7077948_fd=0
   0.1      0      7      0  pid=5636308_fd=10
   0.1      0      6      7  pid=5636308_fd=3
   0.0      0      2      2  pid=9896064_fd=3
   0.0      0      2      0  pid=9896064_fd=10
   0.0      0      0   1938  pid=8912978_fd=8
   0.0      0   1925      0  pid=3342490_fd=7
   0.0      0   1938      0  pid=8912978_fd=5
```




- #MBs：表示文件在间隔时间中的数据传输量，单位是MB。
- #opns：表示文件在间隔时间中的打开次数。
- #rds：表示间隔时间中从该文件读取的IO个数。
- #wrs：表示间隔时间中向该文件写入的IO个数。
- file：表示文件名称。
- volume:inode：其中，volume表示该文件对应逻辑卷的名称或者设备号，inode表示该文件对应文件系统的I节点总数。

HP-UX性能监控

- 适配层性能监控
  - 适配层主要是监控FC HBA卡的性能，可以通过scsimgr命令来查看其状态。需要特别关注的是Outstanding I/Os，表示了当前FC HBA卡的IO并发数。



```
bash-4.1# scsimgr get_stat -D /dev/fcd0

        SCSI STATISTICS FOR CONTROLLER : /dev/fcd0

Generic Statistics:

Illegal events                                  = 0
Ctlr Probe events received                      = 3
BUS Reset events received                       = 0
BUS Reset event failures                        = 0
Target path probe failures                      = 0
Total I/Os processed                            = 1079760
Last time cleared                               = N/A

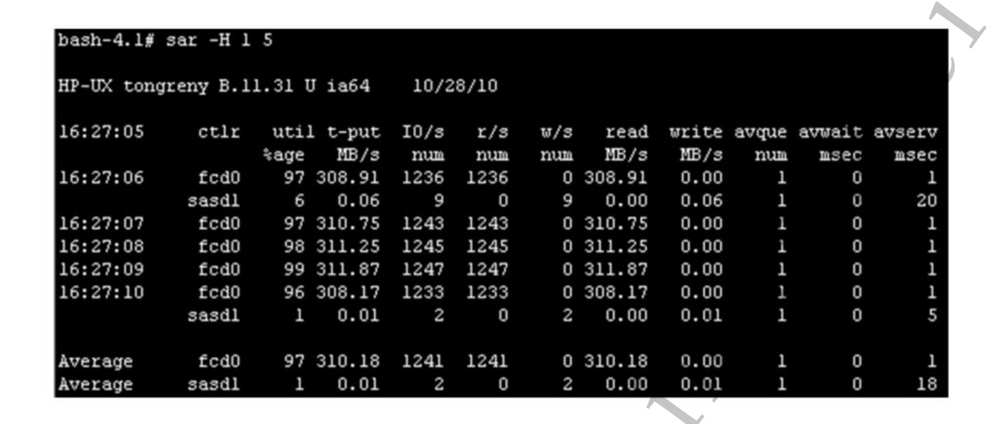
I/F Common Statistics:

Outstanding I/Os                                = 1
Bytes read                                      = 8552910840
Bytes written                                   = 1543946240
Last bus reset time                             = N/A
Target ports connected                          = 1
Bus resets attempted                            = 0
Offline state                                   = 0
Online state                                    = 1
```




HP-UX性能监控

- 适配层性能监控
  - 通过sar命令来查看当前所有处于活动状态的HBA卡的性能。



```
bash-4.1# sar -H 1 5

HP-UX tongreny B.11.31 U ia64    10/28/10

16:27:05      ctlr   util t-put  IO/s   r/s   w/s   read  write avque avwait avserv
                     %age  MB/s   num   num   num   MB/s   MB/s   num   msec   msec
16:27:06      fcd0     97 308.91  1236  1236     0 308.91  0.00     1      0      1
              sasd1     6   0.06     9     0     9   0.00  0.06     1      0     20
16:27:07      fcd0     97 310.75  1243  1243     0 310.75  0.00     1      0      1
16:27:08      fcd0     98 311.25  1245  1245     0 311.25  0.00     1      0      1
16:27:09      fcd0     99 311.87  1247  1247     0 311.87  0.00     1      0      1
16:27:10      fcd0     96 308.17  1233  1233     0 308.17  0.00     1      0      1
              sasd1     1   0.01     2     0     2   0.00  0.01     1      0      5

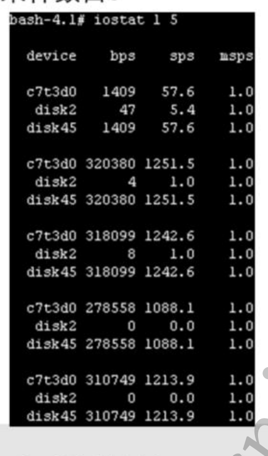
Average       fcd0     97 310.18  1241  1241     0 310.18  0.00     1      0      1
Average       sasd1     1   0.01     2     0     2   0.00  0.01     1      0     18
```




- ctlr：表示HBA卡的逻辑名
- util：表示HBA卡的繁忙程度，单位是百分比
- t-put：表示HBA卡的数据传输率，单位是MB/s
- IO/s：表示HBA卡的IOPS
- r/s：表示HBA卡的读IOPS
- w/s：表示HBA卡的写IOPS
- read：表示HBA卡的读数据传输率，单位是MB/s
- write：表示HBA卡的写数据传输率，单位是MB/s。
- avque：表示HBA卡的当前IO队列深度
- avwait：表示HBA卡的平均IO等待时间，单位是ms
- avserv：表示HBA卡的平均IO服务时间，单位是ms

HP-UX性能监控

- DSF性能监控
  - 通过iostat [interval] [count]来监控DSF性能状况，interval是采样的间隔时间，count是采样数目。



```
bash-4.1# iostat 1 5

  device    bps    sps   msps

  c7t3d0   1409   57.6    1.0
   disk2     47    5.4    1.0
  disk45   1409   57.6    1.0

  c7t3d0 320380 1251.5    1.0
   disk2      4    1.0    1.0
  disk45 320380 1251.5    1.0

  c7t3d0 318099 1242.6    1.0
   disk2      8    1.0    1.0
  disk45 318099 1242.6    1.0

  c7t3d0 278558 1088.1    1.0
   disk2      0    0.0    1.0
  disk45 278558 1088.1    1.0

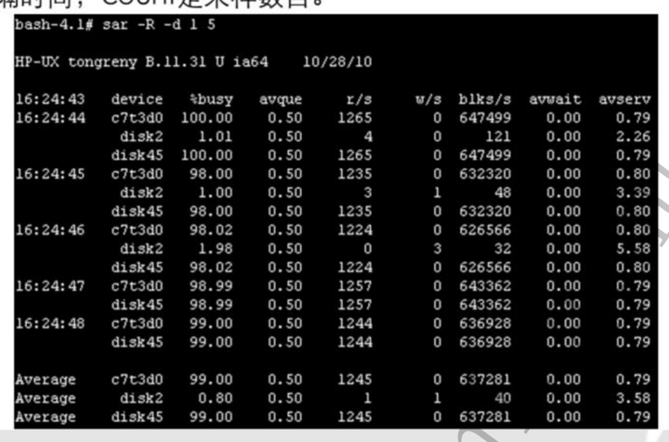
  c7t3d0 310749 1213.9    1.0
   disk2      0    0.0    1.0
  disk45 310749 1213.9    1.0
```

 Page 60 HUAWEI

- bps：表示磁盘的数据传输速率，单位是KB/s。
- sps：表示磁盘的IOPS。
- msps：表示IO的平均响应时间，单位是ms。
- 虽然该命令没有列出平均IO大小，但是可以通过公式计算得到，平均IO大小=bps/sps。

HP-UX性能监控

- DSF性能监控
  - 通过sar -d [interval] [count]来监控DSF性能状况，interval是采样的间隔时间，count是采样数目。



```
bash-4.1# sar -R -d 1 5

HP-UX tongreny B.11.31 U ia64    10/28/10

16:24:43   device   %busy   avque    r/s    w/s  blks/s  avwait  avserv
16:24:44   c7t3d0  100.00    0.50   1265      0  647499    0.00    0.79
            disk2    1.01    0.50      4      0     121    0.00    2.26
           disk45  100.00    0.50   1265      0  647499    0.00    0.79
16:24:45   c7t3d0   98.00    0.50   1235      0  632320    0.00    0.80
            disk2    1.00    0.50      3      1      48    0.00    3.39
           disk45   98.00    0.50   1235      0  632320    0.00    0.80
16:24:46   c7t3d0   98.02    0.50   1224      0  626566    0.00    0.80
            disk2    1.98    0.50      0      3      32    0.00    5.58
           disk45   98.02    0.50   1224      0  626566    0.00    0.80
16:24:47   c7t3d0   98.99    0.50   1257      0  643362    0.00    0.79
           disk45   98.99    0.50   1257      0  643362    0.00    0.79
16:24:48   c7t3d0   99.00    0.50   1244      0  636928    0.00    0.79
           disk45   99.00    0.50   1244      0  636928    0.00    0.79

Average    c7t3d0   99.00    0.50   1245      0  637281    0.00    0.79
Average     disk2    0.80    0.50      1      1      40    0.00    3.58
Average    disk45   99.00    0.50   1245      0  637281    0.00    0.79
```

 Page 61



- %busy：表示磁盘利用率，表明了磁盘的繁忙程度。
- avque：表示主机块设备层IO队列深度。
- r/s：表示磁盘每秒的读IO个数。
- w/s：表示磁盘每秒的写IO个数。
- blks/s：表示磁盘的传输速率，单位是512bytes/s。
- avwait：表示主机块设备层的平均IO等待时间，单位是ms。
- avserv：表示主机块设备层的平均IO服务时间，单位是ms。



目录




配置调优思路
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 63
HUAWEI

应用模块性能分析

业务系统上承载的应用多种多样，按用户下发请求后，其主要表现的IO特性和性能需求，可划分为四大类：

应用场景	业务特征	性能需求
OLTP	小数据块，通常2-8KB 访问位置非常随机 20%-60%写，并发高	高IOPS，低延迟
OLAP	大数据块，通常64-512KB 多路顺序，>90%读	高带宽
多媒体	大数据块，通常32K-1MB 全读或全写，并发高	高带宽、低延迟
虚拟桌面	小数据块，通常<64K 随机访问，>80%读	高IOPS



数据容器性能调优－数据库

调优项	推荐值
表空间	让尽量多的存储资源分担热点区域；合理选取Big/Small File
Cache	采用80%左右的主机内存作为数据库的CACHE使用
数据块	OLTP 4KB或8KB，OLAP 32KB
预读窗口	与ASM/LVM/LUN的条带对齐，建议512K、1MB
索引	删除不必要索引，合理选择B树或位图
分区	大于1亿条记录时分区，合理使用RANGE、LIST、HASH
刷盘进程数	保证无free buffer waits
日志文件	大小适中，建议32-128MB，每实例5个



数据容器性能调优-文件系统

文件系统容器，主要负责处理由上层模块下发的针对文件或目录的操作。

1．选择恰当的文件系统

文件系统分为日志文件系统和非日志文件系统两类.

业务场景	具体业务	适用文件系统
小文件，随机操作	database server，Mail server，小规模电子商务系统，金融系统	Ext3，Reiserfs
大文件，多路顺序读	Video server，	XFS
大文件，多路顺序写	视频监控系统	XFS

服务器CPU个数	适用文件系统
<=8	Ext3 ，Reiserfs
>8	XFS

 Page 66



- 为了数据可靠性，用户一般会选择日志文件系统进行应用业务的搭建。
- 各类文件系统针对某些特定应用进行了专门的性能优化设计，并对主机硬件资源有不同的要求，需根据实际应用选择恰当的文件系统，图表以Linux下的文件系统为例。

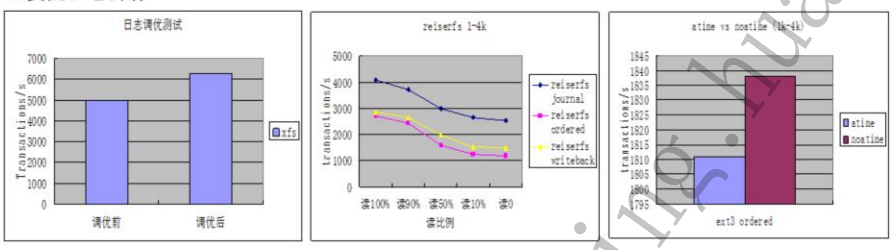
数据容器性能调优-文件系统

2．文件系统参数调整

Block Size调整：同数据应用类似，在OS容许的情况下，建议针对OLTP应用采用4-8K的块大小，针对OLAP及Multimedia应用采用32-64K的块大小

文件碎片整理：文件系统在长时间频繁使用后会出现老化现象，导致文件碎片和空间碎片增加，数据访问时会产生额外的IO开销，同时访问数据也变得更加离散．因此建议根据实际情况，在文件系统使用容量达到60%时进行碎片整理．最佳的碎片整理方式是将所有文件拷贝到另一个相同文件系统上，使元数据与文件数据重新布局．

元数据优化：元数据优化包含三个步骤：元数据与实际数据分离、选择恰当的日志模式、关闭不必要的日志操作



日志区域分离性能对比　　不同日志模式性能对比　　关闭读文件时记录日志性能对比

HUAWEI

数据容器性能调优-文件系统

2．文件系统参数调整

Block Size调整：同数据应用类似，在OS容许的情况下，建议针对OLTP应用采用4-8K的块大小，针对OLAP及Multimedia应用采用32-64K的块大小

文件碎片整理：文件系统在长时间频繁使用后会出现老化现象，导致文件碎片和空间碎片增加，数据访问时会产生额外的IO开销，同时访问数据也变得更加离散．因此建议根据实际情况，在文件系统使用容量达到60%时进行碎片整理．最佳的碎片整理方式是将所有文件拷贝到另一个相同文件系统上，使元数据与文件数据重新布局．

元数据优化：元数据优化包含三个步骤：元数据与实际数据分离、选择恰当的日志模式、关闭不必要的日志操作



操作系统性能调优—内核资源

1

减少多余资源占用:

1.关闭不必要的系统服务与demons

2. 选择恰当的系统启动级别

2

CPU调优:
1.确保各CPU子核负载均衡
2.在处理大量多线程任务时，开启超线程功能

虚拟CPU数	性能提升比率
2	15%-25%
4	1%-13%
8	0%-5%

3

内存调优:

1.根据实际情况，对内存的调度算法、预取窗口、刷盘机制及高低水位进行调整
2.根据实际应用调整各缓冲池页面大小与数量



操作系统性能调优—卷管理模块

- 在创建LVM卷时保证所有LUN：
  - 分条一致
  - 容量相等
  - 磁盘数相等
  - RAID级别一样，并归属不同的存储控制器，
  - 分条大小等于LVM分条单元大小，做到负载均衡。

 Page 70

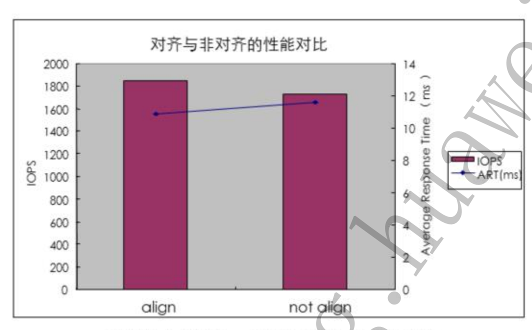


用户下发的请求从数据容器到达卷管理与块设备模块时，已被拆分为多个实际的读写IO请求．这些IO请求有的负责索引节点的查找，有的负责实际数据的访问，还有的则负责日志文件的记录。

卷管理模块能对存储映射的多个LUN进行条带化操作，确保业务负载均匀的分担到不同的LUN上，同时还可设定相应的RAID级别，保证数据可靠性。

操作系统性能调优—块设备模块

- 块设备模块是操作系统中最核心的IO处理模块，它提供了丰富的性能调优参数方便用户根据实际性能需求进行调优。
  - IO对齐
    - IO大小对齐
    - 起始位置对齐
  - 预取窗口调整
  - IO调度策略调整



OLTP应用中进行IO对齐后的性能对比

 Page 71 

块设备模块是操作系统中最核心的IO处理模块，它提供了丰富的性能调优参数方便用户根据实际性能需求进行调优。块设备模块调优主要包括：IO对齐、预取窗口调整、IO调度策略调整。

- IO对齐：
  - IO大小对齐：块设备层可对系统下发到存储的基本IO大小和最大IO大小进行调整．为避免造成业务系统中额外的IO拆分，建议基本IO大小与数据容器设置的基本块大小对齐．最大IO大小与存储及HBA卡可处理的最大IO大小对齐。
  - 起始位置对齐：OS层面，建议为LUN设置一个分条大小的偏移。
- 预取窗口调整
  - 操作系统会根据下发读IO的顺序程度，动态的启停读预取操作，其中某些操作系统的最大预取窗口大小是可设置的。
  - 建议将预取窗口大小与块设备层的最大IO大小对齐，保证在OLAP业务下的读操作命中率得到较大提升，同时不会因为预取IO大小超过系统设置的最大IO大小而导致额外的IO拆分。

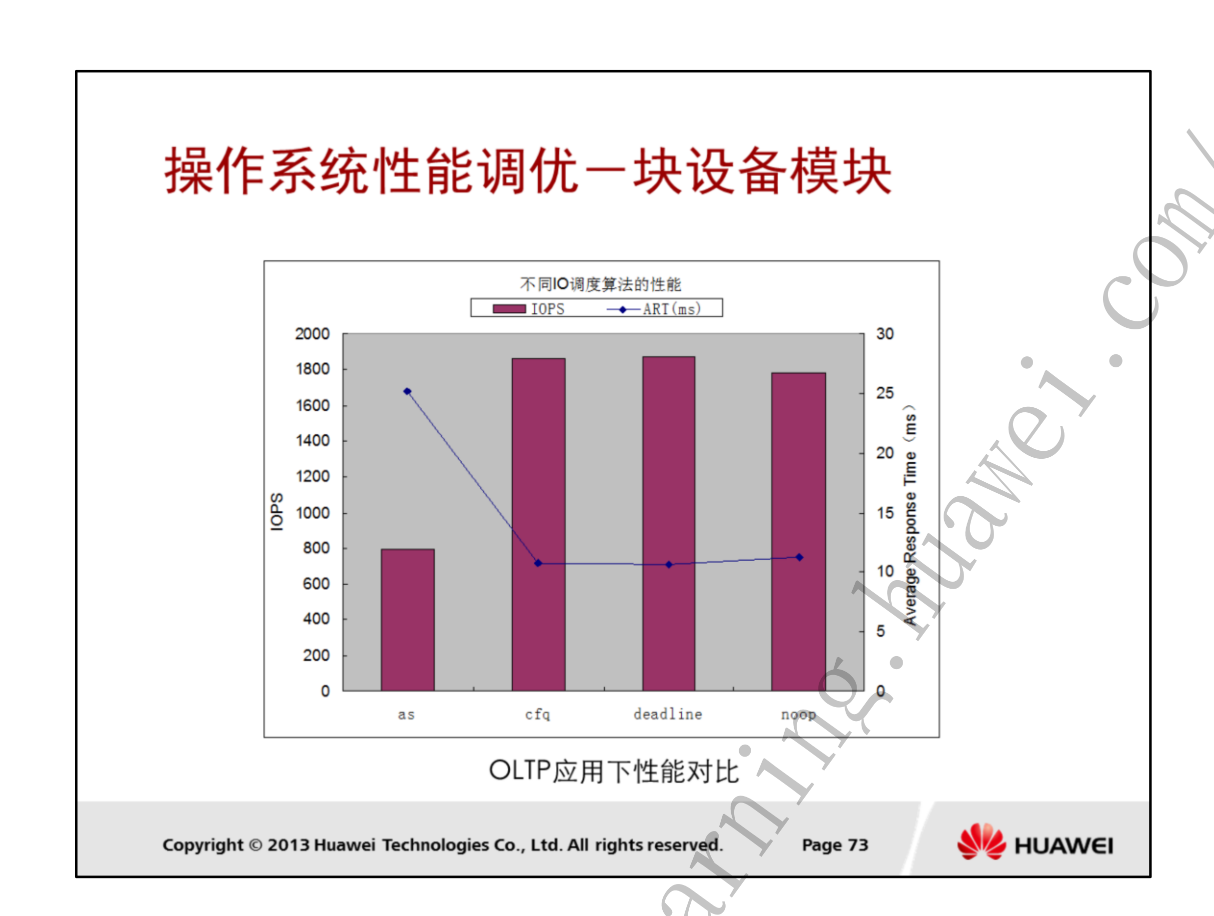
Linux操作系统IO调度策略

调度策略	策略描述	应用场景
noop	除了对相邻IO进行整合外，noop不进行任何调度，以保证IO快速下发	采用SSD硬盘的OLTP、Email、Database等应用
deadline	会对IO进行排序和整合，在IO达到设定时间时立即下发IO	OLTP,OLAP类应用
AS	会对IO进行排序和整合，特别针对读IO会空闲一段时间已保证整合及读命中	适用与IO处理过程中不被中断的应用，不适用于OLTP类应用
CFQ	会对IO进行排序和整合，保证各个进程绝对公平的分享存储资源	适用IPTV、视频监控等多媒体应用

 Page 72



Linux操作系统提供了多种IO调度策略供用户选择，以优化系统的IO性能。请参考上表选择在实际应用中适合的调度策略。


操作系统性能调优—块设备模块
不同IO调度算法的性能
IOPS
ART(ms)
2000
1800
1600
1400
1200
1000
800
600
400
200
0
30
25
20
15
10
5
0
IOPS
Average Response Time（ms）
as
cfq
deadline
noop
OLTP应用下性能对比
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 73
HUAWEI

操作系统性能调优—多路径及HBA卡模块

- HBA卡模块
- HBA卡模块负责IO向存储的下发，需要重点关注如下四个指标：

性能指标	指标描述	8G FC HBA卡性能状况
最大并发数	1．用于描述HBA卡在一个时间片能传输的最大IO数 2．该参数可调，建议调整为最大，避免IO在HBA卡模块出现阻塞	1.单口最大并发为256 2.可通过Execution Throttle参数调整
最大IO大小	HBA卡在不拆分IO的情况下，最大可发送的IO大小	1.通常为 1 MB 2.可通过Frame Size参数调整
最大带宽输出	1．描述HBA卡单口最大的带宽输出 2．需要根据实际应用的存储带宽需求，动态添加HBA卡和网络连接端口数	8G FC HBA卡单口单向带宽为800MB左右
最大IOPS输出	1．描述HBA卡单口最大的IOPS输出 2．需要根据实际应用的存储IOPS需求，动态添加HBA卡和网络连接端口数	8G FC HBA卡单口IOPS输出为50万



操作系统性能调优－多路径及HBA卡模块

- 多路径模块
  - 多路径模块用来控制对存储设备的访问，实现服务器到存储设备之间的路径选择，提高主机与存储设备之间的路径可靠性与性能。
  - 一般包含如下策略:

选路策略	策略说明	适用场景
ROUND_ROBIN	静态负载均衡，轮流在最优路径上下发I/O，以减小单条路径的I/O负荷	适用于I/O负载较小的应用
最小队列深度	动态负载均，优先选择下发I/O过程中未完成I/O数量最少的路径下发I/O	适用于I/O负载较大且对I/O延迟有较高需求的应用，如OLTP应用
最小数据量	动态负载均，优先选择数据量最小的路径下发I/O	适用于I/O负载较大且对传输带宽有较高需求的应用，如OLAP及Multimedia应用



存储系统性能调优概述

CACHE策略选择

Cache策略	推荐配置
读策略	1．顺序业务推荐开启预取，预取窗口根据实际情况调整 2．明确的纯随机随机业务推荐关闭预取 3．IO特性较复杂的业务推荐采用自适应的预取策略
写策略	如无特殊需求，建议采用回写 Cache高低水位建议根据实际情况调整

存储主流RAID类型选择

RAID级别	推荐配置	应用场景
RAID 10	1.推荐采用8盘RAID10 2.建议一个RAID组下只创建一个LUN，分条深度设定为64k～512K	OLTP、Email、Database等
RAID 5	1.推荐采用9盘RAID5 2.建议一个RAID组下只创建一个LUN，分条深度设定为128K	Backup、IPTV、视频监控等
RAID 6	1.推荐采用10盘RAID6 2.建议一个RAID组下只创建一个LUN，分条深度设定为128K	对数据可靠性有较高要求的应用



- 存储并发调整及资源调整根据实际业务情况需考虑对存储各模块并发进行调整，主要调整模块包括：存储前端，CACHE，RAID
- 资源调整主要针对CACHE读写配额，目的是为某些特定测试场景提供更多的CACHE资源，已保证IO整合与调度

存储系统性能调优概述

- 存储与服务器相连网络交换设备调整，保证网络私有：
  - 为避免存储系统与服务器间的网络传输通道成为瓶颈或受到其他业务的干扰，最佳情况下建议采用直连或搭建私有网络将其与服务器连接，并保证网络连接性能与实际业务需求性能匹配.
  - 在网络交换设备有限的情况下，需要在交换机上划分Zone或Vlan，保证两者连接在逻辑上独立



存储系统性能调优—硬盘

不同类型磁盘性能差异

硬盘类型	IOPS/延迟	带宽
SSD	>2000/1ms	100 MB/s
15K HDD	200/10ms	30 MB/s
10K HDD	150/15ms	20 MB/s
7.2K HDD	50/20ms	10 MB/s

传统硬盘内外道性能差异

Cheetah 15K.5 FC HDD

磁盘内外道顺序性能对比

Cheetah 15K.5 FC HDD

磁盘内外道随机性能对比

 Page 78 HUAWEI

1.将热点区域尽量分布在高性能硬盘的外道。

2.对于随机访问,建议在成本容许的情况下，尽量多的使用高性能硬盘来分担负载。

存储系统性能调优－硬盘

OLTP应用硬盘数量估算

由于OLTP应用中,硬盘常常成为性能瓶颈,需要大量的磁盘分担负载,此处给出大致的估算方法作为参考:

1. 通过采集真实应用中，服务器块设备层和存储前端的性能数据，估计系统的最大读\写IOPS输出(对应Physical reads和Physical writes)
2. 通过观测真实应用中存储硬盘侧性能状况，估算磁盘单盘的IOPS输出
3. 根据存储设置的具体RAID级别，估算系统所需硬盘数

RAID10:硬盘总IOPS=Physical Reads+2*physical writes
RAID5:硬盘总IOPS=Physical Reads+4*physical writes
RAID6:硬盘总IOPS=Physical Reads+6*physical writes

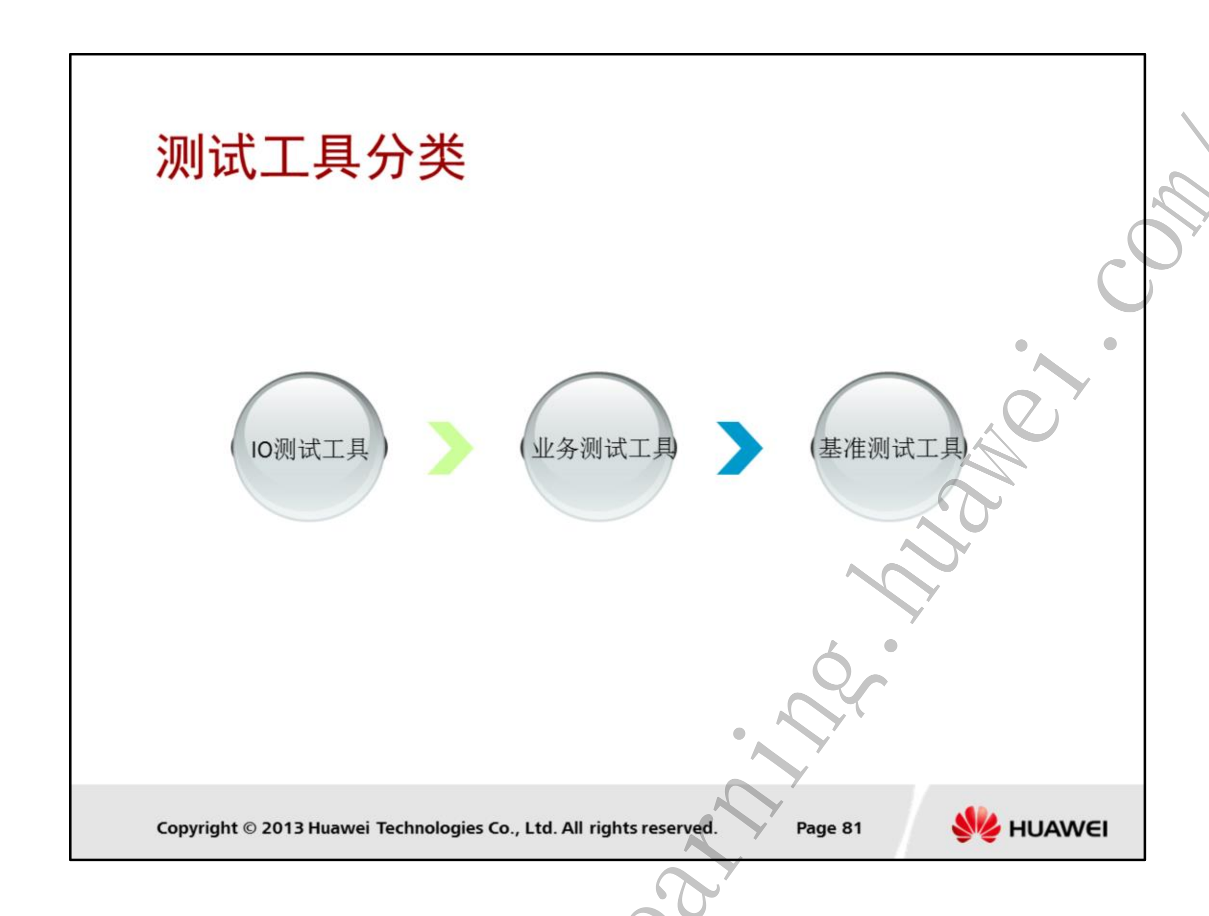




目录

1. 性能调优概述
2. 性能指标
3. 影响性能的关键因素及技术
4. 性能诊断和调优
5. **性能测试工具**
6. SAN存储系统常见性能故障排除





- IO测试工具：提供IO负载，主要用于最基本的性能压力测试，通常绕过上层结构直接对裸设备进行测试，可以通过一些测试设置大体模拟某些实际应用场景。
- 业务测试工具：模拟实际业务场景进行针对该场景下存储性能测试的工具，这些工具通常是针对某些特定的上层系统在实际业务中的应用来进行测试的，这样的测试结果更具参考价值。
- 基准测试工具：存储性能三大基准组织提供的针对其各种性能基准的测试工具。本着公平、公正、有可比性的原则，制定出的性能测试工具一般是与存储最广泛的应用场景强相关的，具有很高的度量和对比价值。

IO性能测试

测试工具	适用范围	缺陷
Iometer	适用操作系统: Windows,Linux,Solaris,Netware,Mac OSX 适用场景: 提供非常全面的IO特性配置参数,可作为IO压力测试工具，能在IO层面模拟一些简单的业务场景,支持联机测试	1.Linux存在异步IO压力不足的问题 2.无法进行数据一致性验证 3.联机测试中容易出错
XDD	适用操作系统: Windows,Linux,Solaris,Mac OSX,AIX,HP UNIX,IRIX 适用场景: 可作为IO压力测试工具,目前在AIX下使用最多	1.存在一些兼容性bug,在某些主机或操作系统上运行会报错 2.无法设置测试时间 3.测试时工具本身有一些性能波动 4.不支持联机测试
IORate	适用操作系统: Linux,Solaris,AIX,HP UNIX 适用场景: 可作为IO压力测试工具，可以模拟相同IO特性的不同压力等级	1.IO特性配置参数不够丰富 2.不支持在Windows平台上测试 3.不支持联机测试



业务场景测试

测试工具	适用范围	缺陷
IOZone	适用操作系统：Windows,Linux 适用场景：测试不同类型文件系统在Read, write, re-read, re-write, read backwards, read strided, fread, fwrite, random read, pread ,mmap, aio_read, aio_write 时的性能状况	1.只能作为测试存储产品与特定主机和文件系统搭配时的性能测试工具，基本不具备场景模拟功能 2.不具备对文件创建＼删除＼遍历和对目录操作的能力 3.无法联机测试；不具备数据一致性校验能力
Postmark	适用操作系统：Windows,Linux,Solaris 适用场景：Email Server，小文件应用场景和文件系统老化测试	1.单线程程序，压力不充分 2.通过随机函数实现文件的读/写及创建/删除，操作不可控，不具备目录操作能力 3.只适用于小文件操作场景的模拟 4.无法联机测试；不具备数据一致性校验能力
Boine++	适用操作系统：Linux和UNIX相关平台 适用场景：测试文件各种读写方式性能，测试大量文件创建删除	类似IOZone，但支持的文件操作和可控操作比IOZone更少．一般不使用
Orion	Oracle性能测试基准工具，针对Oracle数据的OLTP，Backup和OLTP+Backup混合业务进行模拟．无须安装oracle数据库即可用于测试存储在oracle下的性能	1.无法联机测试 2.无法模拟Oracle各特定区域的IO特性

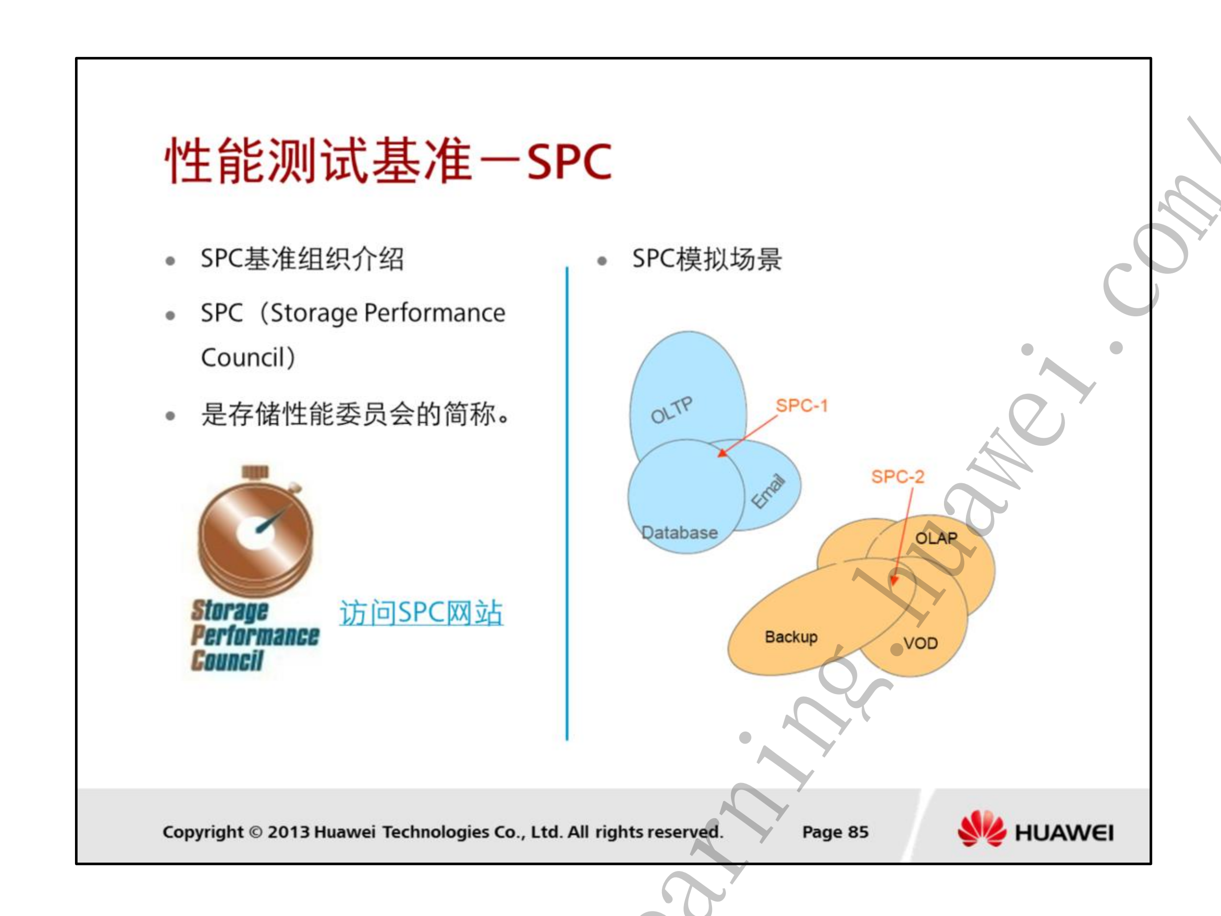


- IOZone测试对象：文件系统，计算机CPU，内存和存储设备
- Postmark测试对象：文件系统，计算机CPU，内存和存储设备
- Boine++测试对象：主要是文件系统和存储设备

性能测试基准

测试工具	工具描述
SPC-1	SPC-1性能基准配套测试工具用于评估存储产品在OLTP\Email\Database三种应用场景下的性能状况
SPC-2	SPC-2性能基准配套测试工具用于评估存储产品在Backup\VOD（video on demand）\OLAP三种应用场景下的性能状况
SPC-3	SPC-3BR: 基于存储管理软件的性能测试基准，度量数据备份和恢复的性能，该基准暂未正式公布 SPC-3ILM: 基于存储管理软件的性能测试基准，度量多存储系统节点所组成的大型存储解决方案的性能．目前该基准还处于构想阶段
TPC-C	构造一个典型的批发商应用模型，验证由服务器、数据库、存储所搭建的一整套业务系统在OLTP业务下的性能． TPC-C规范只提供了规范和工具的实现方式，并未提供实际测试工具
SPEC-SFS2008	验证服务器和存储搭建的业务系统在NAS业务下的性能.支持CIFS和NFS



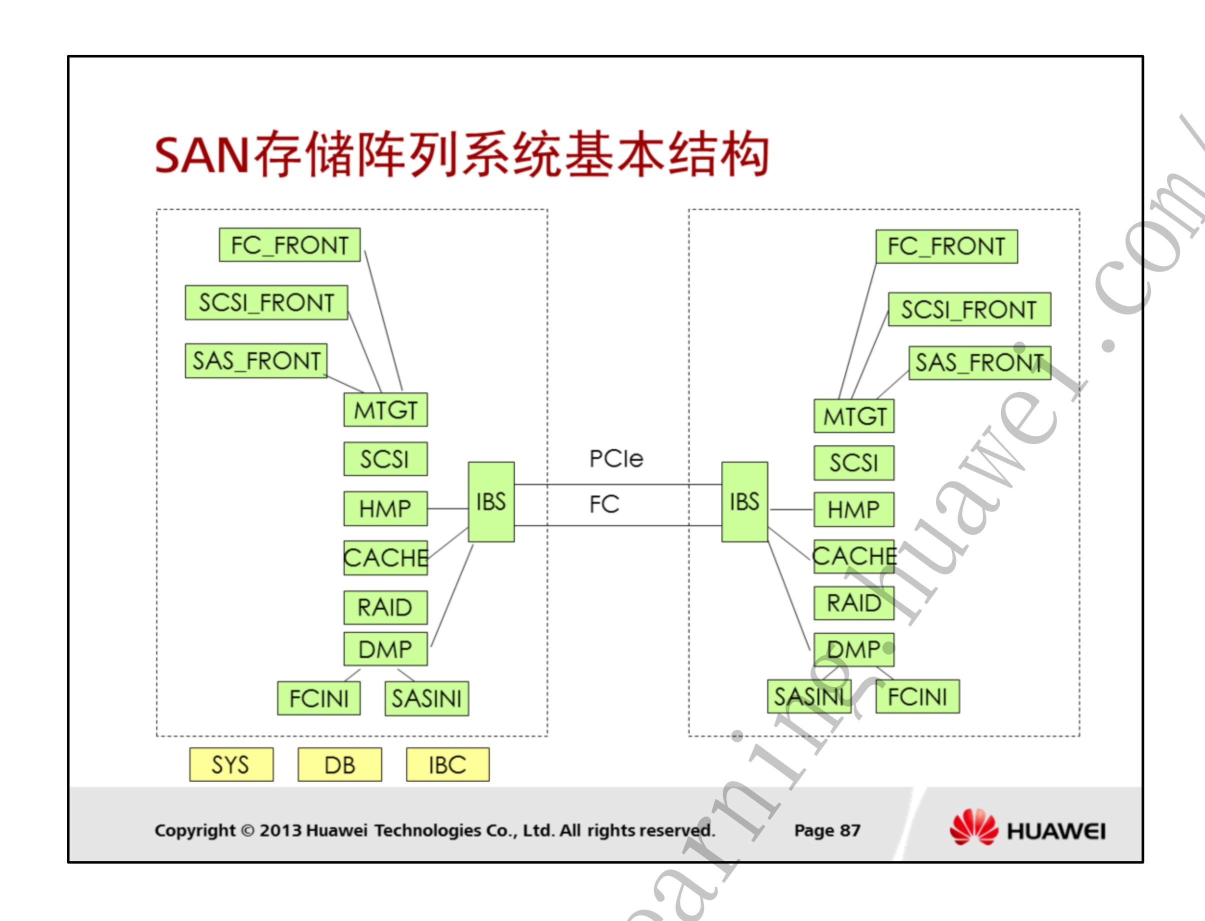


SPC自身是一个非营利组织，它的建立就是为了定义、标准化以及促进存储子系统的基准，同时向计算机厂商和用户发布客观、权威、公正的性能数据，为客户提供参照，为存储厂商提供交流平台携手提升存储性能。它公布的性能数据得到了各大存储厂商和客户的广泛认可。

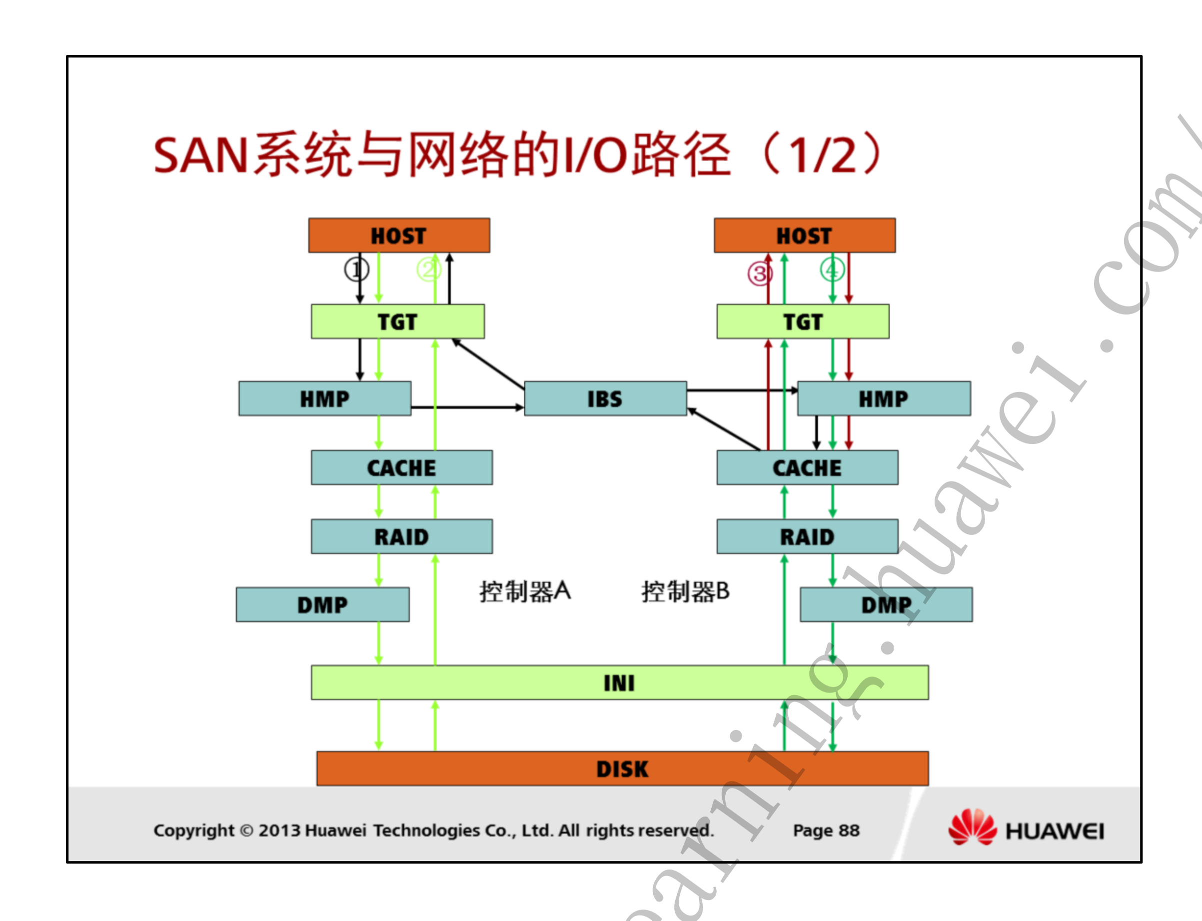


目录

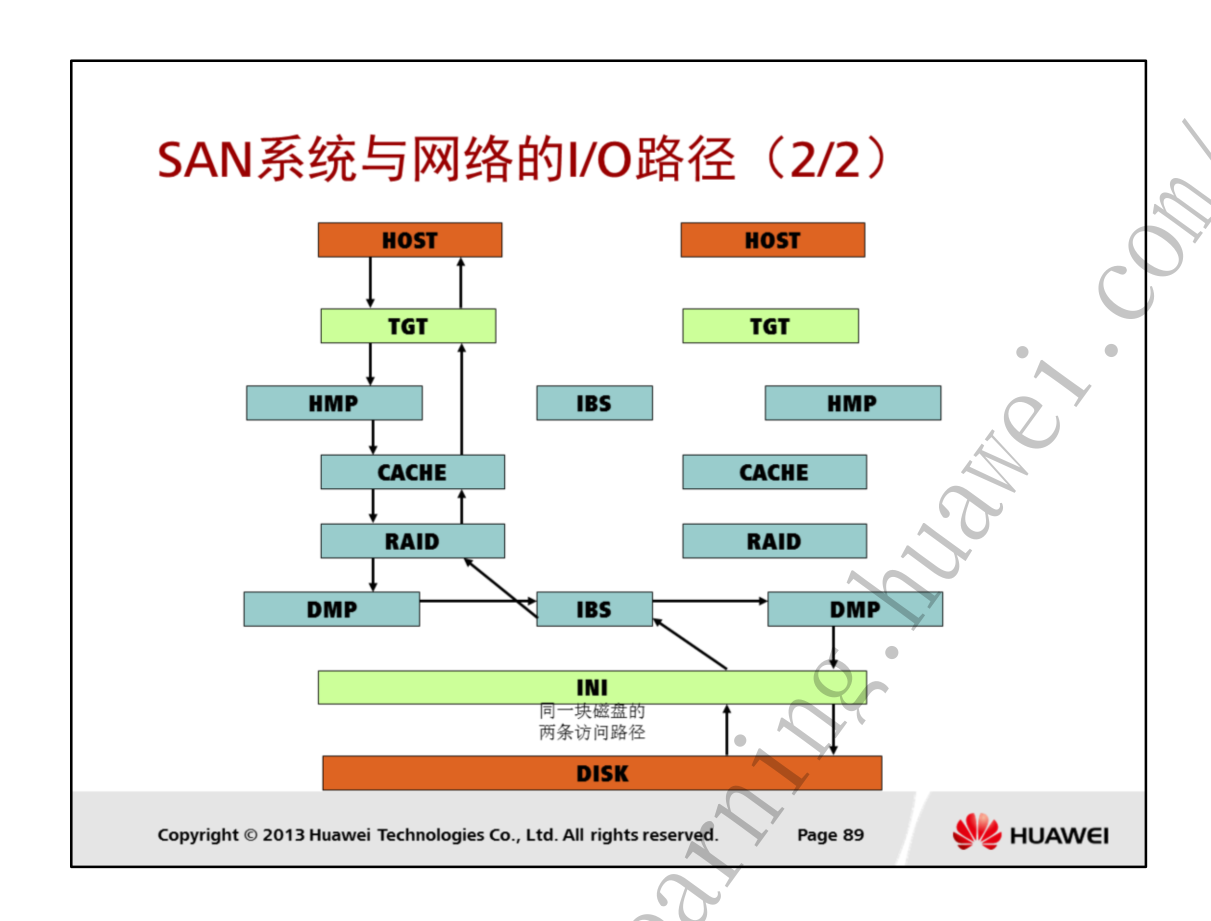




- XX_front：具体处理主机接口协议的传输层，例如：FC、SCSI（iSCSI）、SAS
- MTGT：目标器中层，负责SCSI命令状态管理和前后调度
- SCSI：负责SCSI命令的解析和处理
- HMP：多路径控制模块与主机多路径交互，负责把读写命令发送到合适的控制器处理
- CACHE：负责cache管理，例如回写、透写、预取等
- RAID: RAID管理和数据组织，例如5/10/0 重构、copyback
- DMP：当控制器访盘失败后，启动对端控制器访问硬盘
- FCINI：FC 协议启动器
- SASINI：SAS 协议启动器
- IBS：负责控制器间通讯管理（主要是PCIe）
- SYS：系统管理，控制各业务模块的启动、停止等
- DB：管理和保存系统的配置数据
- IBC：板间通讯模块，负责管理命令的通信



- TGT：目标器中层，负责SCSI命令状态管理和前后调度
- HMP：多路径控制模块与主机多路径交互，负责把读写命令发送到合适的控制器处理
- IBS：负责控制器间通讯管理（主要是PCIe）
- CACHE：负责cache管理，例如回写、透写、预取等
- RAID: RAID管理和数据组织，例如5/10/0 重构、copyback
- DMP：当控制器访盘失败后，启动对端控制器访问硬盘
- INI：协议启动器
- ①表示回写镜像的IO路径。
- ②表示透写的IO路径。
- ④ 表示读未命中，从硬盘中读数据
- ③表示读命中，直接从cache中读到数据



当本端RAID模块到DISK的路径出现异常时，就会通过DMP模块，走IBS通道，到对端去处理IO，这时IO路径更长，性能会下降。

存储侧性能问题排查思路

读写性能问题
- LUN透写
  - 配置为透写
  - 阵列故障
    - 电池故障
    - 电源故障
    - 单控运行
- 链路问题
  - iSCSI链路
  - FC链路
- 其它问题
  - 存在慢盘
  - LUN是否正在格式化
  - RAID组，LUN配置不合理
  - 小数据随机IO
  - 硬盘双端访问

存储侧性能问题排查方法与步骤

- 排查步骤
  - 查看LUN状态；
  - 查看电池状态；
  - 查看电源状态；
  - 查看控制器状态或者观察控制器指示灯状态；
  - 查看网口的速率；
  - 主机上使用大数据包ping命令，如windows下：ping [ip] -l 20000 -t;
  - 调试模式下执行iostat -x 1查看硬盘svctm是否超过50ms或者某个硬盘svctm明显高于其他硬盘；
  - 查看同一个RAID组中LUN是否归属于不同的控制器；
  - 查看LUN的工作控制器是否和接收IO的控制器相同。



存储侧性能问题排查案例

- 描述问题
  - 某IPTV局点客户反馈，偶尔出现黑屏情况。
- 分析原因
  - 系统状态正常，LUN运行策略为回写镜像，链路正常，无误码，无闪断。使用iostat －x 1观察，发现在业务量大时，有一个盘的服务时间（svctm）明显比其他盘高很多，超过50ms，该盘为慢盘。
- 解决步骤
  - 将该硬盘离线，并更换该硬盘。
- 验证恢复
  - 更换慢盘后，没有再出现黑屏现象。





目录



网络层性能问题排查思路

- 网络层性能问题
  - 网络带宽
    - 测试iSCSI组网连通性和带宽
    - 测试FC组网连通性和带宽
  - 网络路径
    - 使用uptool查看路径连接情况
    - 调整多路径选路算法
    - 集群环境是否关闭failover功能

网络层性能问题排查方法与步骤

- 排查网络带宽的常见方式有如下几种：
  - 针对iSCSI的组网方式，可以使用ping包的命令，从主机直接ping阵列的业务网口，查看网络延迟和丢包现象。
  - 针对FC组网方式，可以先通过命令showfreeport 查看主机端口是否连通，然后使用showfps查看端口速率。通过确认端口速率排除端口速率协商过低的问题。
  - 针对FC组网方式，可以在ISM查看端口误码信息，排除网络传输误码导致的性能问题。



网络层性能问题排查方法与步骤

- 排查网络路径的常见方式有如下几种：
  - 通过使用多路径命令upadm －S 查看主机到阵列路径数以及连通状况。
  - 针对主机到阵列多条路径的情况，可以通过调整主机上多路径的选路算法来实现存储系统的性能提升。
  - 集群的一种场景，主机没有关闭多路径的failover功能，导致路径切换影响读写性能。



网络层性能问题排查案例

- 描述问题
  - 客户反馈将LUN映射给集群主机后，测试性能下降较多。
- 分析原因
  - 查看存储系统正常， LUN运行策略为回写镜像，但主机不断下发LUN路径切换命令。
  - 排查发现其中一台主机的A控链路故障后，下发了切LUN命令至B控，而对另外一台主机，A控才是LUN真正的归属和工作控制器，同样也下发切换LUN命令至B控。
- 解决步骤

  1、修复故障的链路

  2、关闭多路径的failover功能，解决该问题
- 验证恢复

  1、链路修复后，读写性能能够恢复正常

  2、直接关闭多路径的failover功能，读写性能也能够恢复正常





目录



主机侧性能问题排查思路

- 主机侧性能问题
  - HBA卡参数设置
  - Windows主机
    - 通过性能监测工具查看数据读写性能
  - Linux主机
    - 使用sar命令查看CPU，内存使用情况
    - 使用iostat命令查看存储使用情况

SAN性能影响因素－主机侧

- HBA卡
  - 最大单个请求大小
  - 最大并发请求数
  - 选择合适的HBA卡驱动

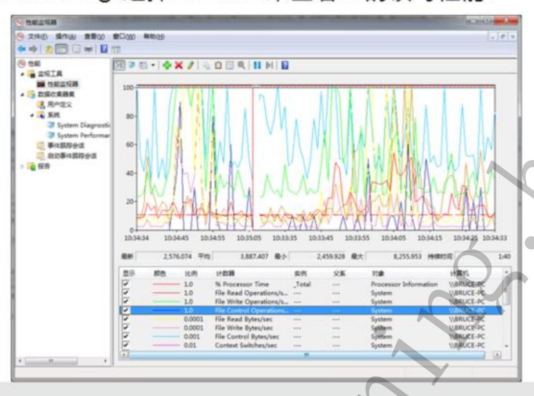
Page 100



- 最大单个请求大小：主机HBA卡通过设置驱动参数来限定下发请求数据的最大值。建议将主机HBA卡最大单个大小请求设置不超过1MB；设置方法参考HBA卡驱动手册。
- 最大并发请求数：主机HBA每个端口有最大请求并发数限制。建议设置在256－512之间。最大请求并发数查看和设置方法参考HBA卡驱动手册。
- 选择合适的HBA卡驱动：比如说将一个1MB的请求拆分为一个1020个扇区和8个扇区的两个请求，导致数据不是满分条下发，影响测试性能。此时可以选择不拆分IO的HBA卡驱动。

主机侧性能问题排查方法

- Windows主机性能查询方法
  - 在windows系统下查看性能，通常会先收集Performance monitor信息，来
  - 首先确认IO性能的状态，运行方法为，开始菜单，运行 “Perfmon” ，可以
  - 新建一个Counter Logs选择Counter来查看IO的读写性能。



 Page 101 

- 通常的Disk性能指标有:
  - Logical Disk\Avg Disk Sec/Transfer： < 10ms 比较好;
  - PhysicalDisk\Avg. Disk Read Queue Length： 这个值小于2比较好;
  - PhysicalDisk\Avg. Disk sec/Read: <20 ms 比较好;
  - PhysicalDisk\Avg. Disk Queue Length ： Should not be higher than the number of spindles plus two.;
  - PhysicalDisk\Avg. Disk sec/Transfer。
- 注意，以上任何参数都不是固定的，需要结合起来看。

主机侧性能问题排查方法

- 通常的Disk性能指标有：

  Logical Disk\Avg Disk Sec/Transfer： < 10ms 比较好；

  PhysicalDisk\Avg. Disk Read Queue Length： 这个值小于2比较好；

  PhysicalDisk\Avg. Disk sec/Read：<20 ms 比较好；

  PhysicalDisk\Avg. Disk Queue Length ：Should not be higher than the number of spindles plus two.；

  PhysicalDisk\Avg. Disk sec/Transfer。

- 注意，以上任何参数都不是固定的，需要结合起来看。



主机侧性能问题排查方法

- Linux主机性能查询方法

Linux系统资源监控命令主要用于监控CPU利用情况和内存利用情况，主要使用命令如下：

sar [options] [-o file] t [n]

在命令行中，n 和t 两个参数组合起来定义采样间隔和次数，t为采样间隔，是必须有的参数，n为采样次数，是可选的，默认值是1，-o file,表示将命令结果以二进制格式存放在文件中，file为文件名。

```
OceanStor:~ # sar -u 3 3
Linux 2.6.5-7.244-smp (OceanStor)        06/10/08

18:11:25          CPU     %user     %nice   %system   %iowait     %idle
18:11:28          all      0.00      0.00      0.50      0.00     99.50
18:11:31          all     10.83      0.00     10.83      0.00     78.33
18:11:34          all      0.17      0.00      0.50      0.00     99.33
Average:          all      3.67      0.00      3.95      0.00     92.38
```

 Page 103 HUAWEI

- CPU：表示机器内所有的CPU；
- %user：表示CPU的利用率；
- %nice：表示CPU在用户层优先级的百分比，0表示正常；
- %system：表示当系统运行时，在用户应用层上所占用的CPU百分比；
- %iowait：表示请求硬盘I/0数据流出时，所占用CPU的百分比；
- %idle：表示空闲CPU百分比，值越大系统负载越低。

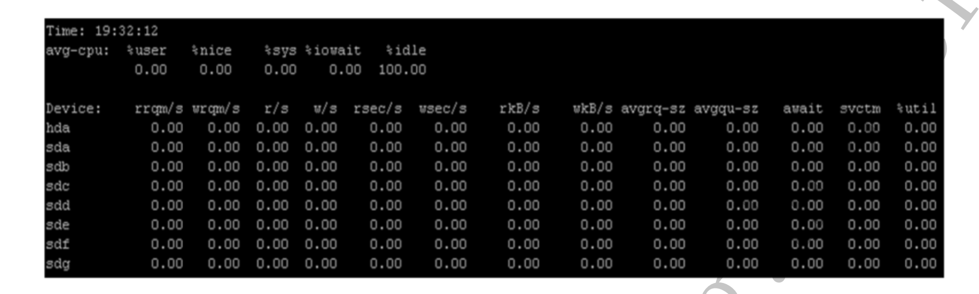


其中await，svctm和%util三个参数比较重要，表示当前硬盘的工作状态：

- await：平均每次设备I/O操作的等待时间（毫秒），即delta(ruse+wuse)/delta(rio+wio)；
- svctm：平均每次设备I/O操作的服务时间（毫秒），即delta(use)/delta(rio+wio)；
- %util：一秒中有百分之多少的时间用于 I/O 操作，或者说一秒中有多少时间 I/O 队列是非空的，即
- delta(use)/s/1000（因为use的单位为毫秒）。

主机侧性能问题排查方法

- iostat命令
  - iostat -x -t 2(打印间隔时间)
  - iostat命令用于显示当前CPU利用状态和各磁盘的使用情况，显示结果如下：



```
Time: 19:32:12
avg-cpu:  %user   %nice    %sys %iowait   %idle
           0.00    0.00    0.00    0.00  100.00

Device:    rrqm/s wrqm/s  r/s  w/s  rsec/s  wsec/s    rkB/s    wkB/s avgrq-sz avgqu-sz   await  svctm  %util
hda          0.00   0.00 0.00 0.00    0.00    0.00     0.00     0.00     0.00     0.00    0.00   0.00   0.00
sda          0.00   0.00 0.00 0.00    0.00    0.00     0.00     0.00     0.00     0.00    0.00   0.00   0.00
sdb          0.00   0.00 0.00 0.00    0.00    0.00     0.00     0.00     0.00     0.00    0.00   0.00   0.00
sdc          0.00   0.00 0.00 0.00    0.00    0.00     0.00     0.00     0.00     0.00    0.00   0.00   0.00
sdd          0.00   0.00 0.00 0.00    0.00    0.00     0.00     0.00     0.00     0.00    0.00   0.00   0.00
sde          0.00   0.00 0.00 0.00    0.00    0.00     0.00     0.00     0.00     0.00    0.00   0.00   0.00
sdf          0.00   0.00 0.00 0.00    0.00    0.00     0.00     0.00     0.00     0.00    0.00   0.00   0.00
sdg          0.00   0.00 0.00 0.00    0.00    0.00     0.00     0.00     0.00     0.00    0.00   0.00   0.00
```

 Page 105



Linux下主机侧性能问题排查常用命令：

- uptime 命令的显示结果包括服务器已经运行了多长时间，有多少登陆用户和对服务器性能的总体评估（load average）。
- top命令显示了实际CPU使用情况，默认情况下，它显示了服务器上占用CPU的任务信息并且每5秒钟刷新一次。你可以通过多种方式分类它们，包括PID、时间和内存使用情况。
- iostat是sysstat包的一部分。iostat显示自系统启动后的平均CPU时间（与uptime类似），它也可以显示磁盘子系统的使用情况，iostat可以用来监测CPU利用率和磁盘利用率。
- vmstat命令提供了对进程、内存、页面I/O块和CPU等信息的监控，vmstat可以显示检测结果的平均值或者取样值，取样模式可以提供一个取样时间段内不同频率的监测结果。
- free命令显示系统的所有内存的使用情况，包括空闲内存、被使用的内存和交换内存空间。Free命令显示也包括一些内核使用的缓存和缓冲区的信息。



思考题

1. 哪些存储配置会对性能产生影响?
2. 华为存储有哪些功能用来提升客户体验?





总结

- 性能指标
- 影响性能的关键因素及技术
- 性能诊断和调优
- 性能测试工具和方法
- SAN存储系统常见性能故障排除





习题

- 判断题
  1. 在顺序写的应用场景下,对于RAID10来说,磁盘数量越多越好（T or F)
- 单选题
  1. Cache中单个chunk最大可以是多大?（　　）
     A. 64KB
     B. 256KB
     C. 512KB
     D. 1MB



- 习题答案:
  - 判断题：1.T
  - 单选题：1.D

Thank you

www.huawei.com

HC120920007 统一存储方案规划与设计




HC120920007
统一存储系统方案规划与设计
www.huawei.com
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
HUAWEI



目标

- 学完本课程后，您将能够:
  - 熟悉统一存储系统规划的目标
  - 熟悉统一存储系统规划的步骤和方法





目录





- 目标：满足客户的需求

业务需求	IT应用需求
✓ 兼容性	✓ 资金预算
✓ 性能	✓ 人力预算
✓ 容量	✓ 可管理性
✓ 可靠性	✓ 资源利用效率
✓ 可扩展性	✓ 节能减排
✓ 灾备需求	
✓可服务性	

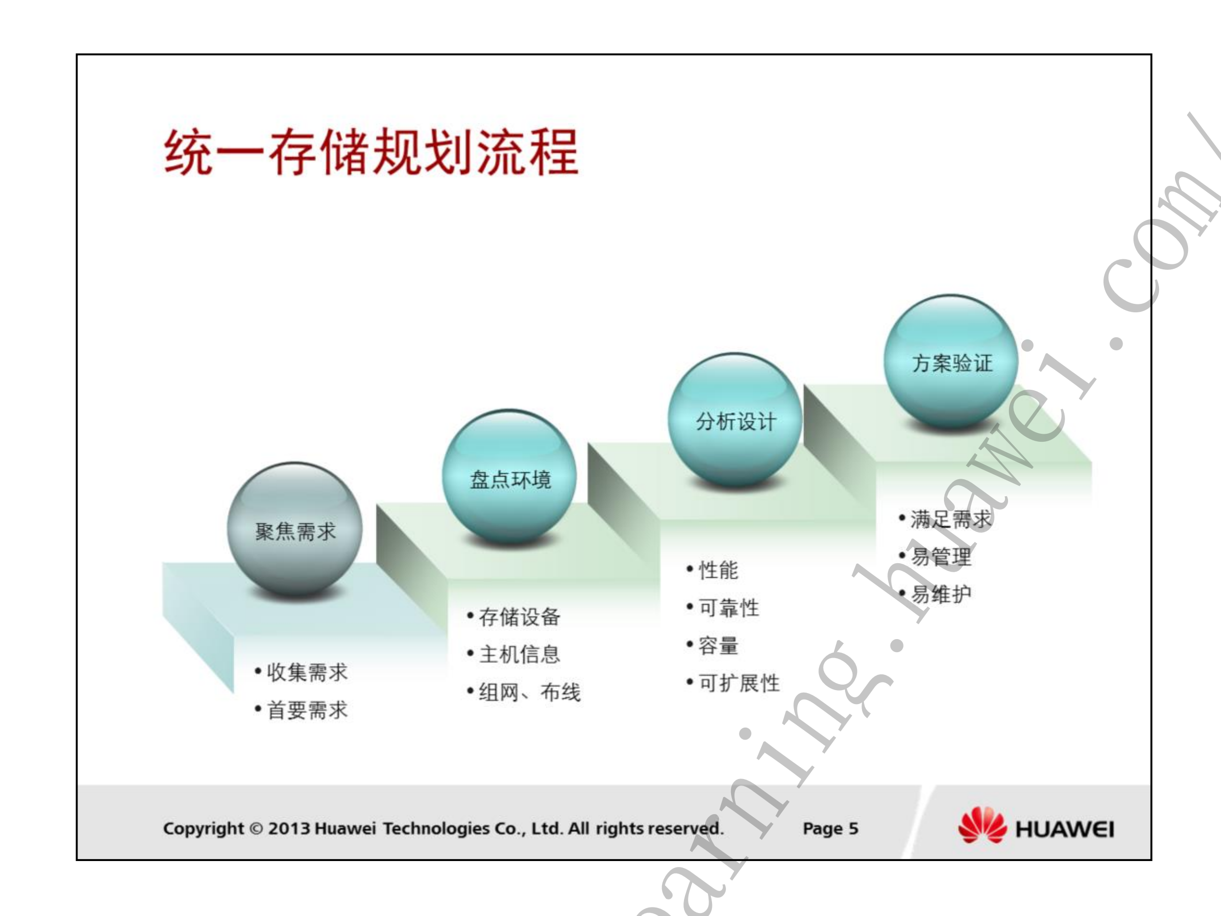
 Page 3 

- 满足客户业务需求：
  - 兼容性，包括OS兼容性、软硬件兼容性，此处兼容性作为选择SAN组件的基础，必须在规划前弄清楚各SAN存储兼容性列表。
  - 性能需求，包括要害业务应用的性能峰值带宽、所有业务系统的性能峰值带宽、各业务系统响应时间、延迟。
  - 容量需求：现有业务存储容量需求、业务增长容量需求（月增长量）。
  - 可靠性，包括存储可靠性、网络可靠性、主机可靠性。
  - 备份与容灾需求，紧急情况下，需要什么应用程序来迅速、有效切换到另一个备用数据中心？是否有备用数据中心？假如有，与主用数据中心的距离有多远？需要在峰值生产时间克隆或快照任务要害型数据
  - 可用性：是否存在计划外停机、不中断业务情况下需新增哪些存储资源
  - 可服务性：问题诊断与故障排除。
  - 业务扩展需求，如存储需求比服务器需求快多少、来年预计存储空间扩容、来年预计新增业务应用（ERP、CRM、MIS等）、未来N年存储规划等。
- 满足客户IT应用需求：主要包括预算与投入、资金投入、人力投入、时间投入、可管理性、简单管理、融易存储，以及资源利用率、节能减排等。

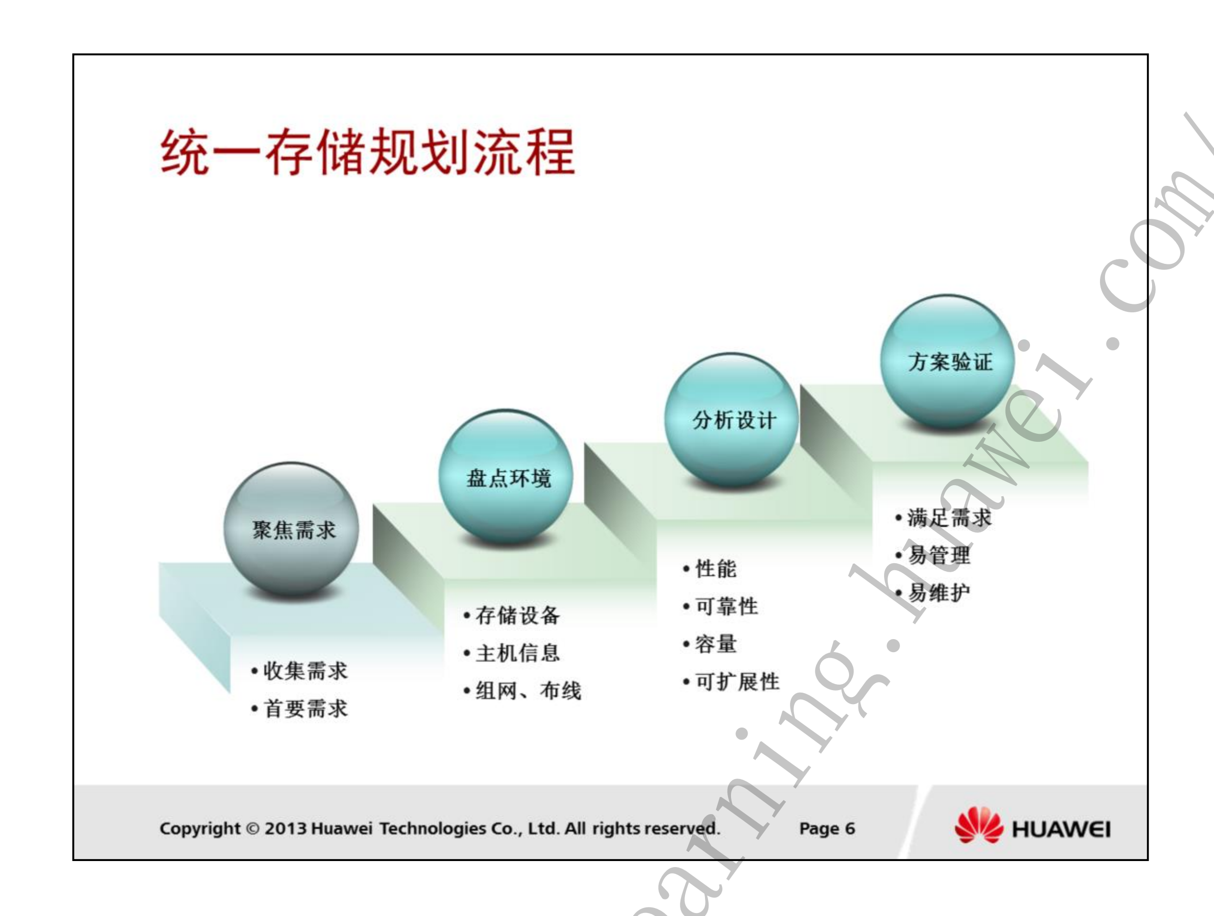


SAN规划总的原则：需求主导，统筹规划，具体在规划SAN时，应依据下面四个具体原则：

- 单一需求原则：业务需求驱动技术选择，最成功的SAN规划与部署来源于单一需求。对比客户的需求列表，明确哪个需求是最至关重要的，然后集中精力围绕该需求进行设计、验证与部署。单一需求不是仅仅关注一个需求，而是以最核心的业务需求为重点，以其他业务需求为补充。
- 简单设计原则：通过对SAN管理单元进行设计，管理工作简化、出错的机率减少，可用性也由此得到提高。经仔细构思和实施的SAN设计可便于随时增加新的服务器和存储目标，同时用户不必头疼于复杂的Fabric架构路径。围绕多个SAN管理单元和SAN路由而设计的大型数据中心存储网络更易于扩展，使得管理员通过可靠且一致的SAN设计战略即能满足其企业不断增长的业务需求。
- 开放性与标准化：设计方案中所采用的技术和选用的产品都必须是标准的、业界公认的主流，而且必须满足开放性的要求。
- 经济性原则：设计方案不但要考虑采用技术的先进、可靠，而且还必须考虑用户的经济负担。因此，设计方案必须具备很高的性能价格比；盘点与分析现网中哪些SAN组件仍然能满足客户的业务需求，并且能够连入SAN环境，如交换机、路由器、网线、光纤等，做到尽量使用客户存在的设备，减少在该方面的投入。



- 聚焦客户需求:
  - 依据SAN规划目标介绍的需求列表，逐条与客户确认各条需求的指标。
  - 对照业务需求列表，确定哪条需求是至关重要的，然后围绕该需求进行规划、设计与验证。
  - 成功运营SAN后，利用运营结构来展示投资回报情况，并可部署更多SAN满足业务应用或者扩展SAN。
- 盘点环境
  - 对比已存在SAN存储，需增加支持哪些业务需求.
  - 盘点现有组件，明确哪些可继续使用，如主机信息、存储设备信息、网络设备信息、盘点组网与环境，确定组网限制与布线等。
  - 盘点组件信息，包括主机操作系统、HBA个数以及驱动程序（包括驱动程序的版本号）、所支持连接的类型（环路或者光纤通道）、应用清单、初步和预期存储要求 、尺寸以及重量 ，存储设备包括品牌、型号以及固件版本 、所支持连接的类型（环路或者光纤通道、可支持的主机数量 、端口以及每个端口可支持的主机数量、容量（已用容量以及空闲容量） 、光纤通道接口的数量以及类型 、以太网接口的数量以及类型等。



- 分析设计
  - 分析库存清单、明确可重用组件、选择新组件应考虑哪些物理需求
  - 从性能、可靠性、容量与可扩展性等方面规划与设计SAN。
- 方案验证
  - 验证前面规划的SAN方案能否满足需求。

统一存储系统规划

主机规划	网络层规划	存储层规划
• 主机配置 • 文件系统 • 业务软件	• IP SAN与FC SAN选择 • SAN网络拓扑 • VLAN/zone划分	• 控制器性能 • RAID组规划 • LUN归属 • 分条深度 • Cache预取算法 • Cache写策略 • Cache高低水位 • 磁盘选择





目录



主机操作系统规划

操作系统	最新版本	应用场景
AIX	7	IBM小型机、IBM刀片服务器
solaris	11	SUN小型机、X86架构服务器
HP-UX	11iV3	HP小型机（PA-RISC、Itanium）
windows	2012	刀片、机架服务器
linux	suse12/rhel7	刀片、机架服务器
vSphere	5	刀片、机架服务器



- 需要根据客户的业务的关键程度、性能需求等因素来考虑使用的操作系统类型和相应的硬件类型对于关键类型的应用可以采用小型机来作为主机平台。
- 虚拟机目前正在成为主机部署的热点。

主机文件系统规划

OS	文件系统	最大卷容量	文件最大尺寸
Windows	FAT32	8TB	4GB
	NTFS	16EB	2TB
Linux	Ext2	32TB	2TB
	Ext3	32TB	2TB
AIX	JFS2	64PB	1PB
HP-UX	HFS	128GB	/
	JFS3.5	2TB/32TB	/

 Page 10



- 文件系统是操作系统用于明确磁盘或分区上文件的方法和数据结构，即在磁盘上组织文件的方法。
- 裸盘在分区和创建文件系统之后容量是有损失的。
- 文件系统不同，其文件系统建立后导致的容量损失也不尽相同。

双机和集群技术介绍

高可用集群（High Availability）

致力于提供高可靠服务

– 高可靠性集群

– 负载均衡集群

高性能集群（High Performance）

致力于提供单个计算机所不能提供的强大的计算能力

– 高吞吐计算集群

– 分布计算集群



- 集群又称为机群或群集。始于上世纪70年代，用于科学计算，随着Linux集群的出现，集群才开始推广。
- 高可用(High Availability)集群,简称HA集群。这类集群致力于提供高度可靠的服务。高可靠集群是高可用集群的一种，一般具有如下功能：
  - 硬件全冗余：心跳线、业务网口、存储连接等；
  - 故障自动切换：集群软件自动监控故障(包括软、硬件故障)、提供切换策略（包括本地切换与服务器间切换）。
  - 适用情况：高可靠性要求；允许业务短暂中断；负载较小。
- 高可靠集群特点：
  - 不需操作者干涉的情况下自动故障恢复；当然，对失败节点的故障定位及修复，还是需要专门技术人员进行操作。
  - 硬件上尽量避免单点故障，各部件冗余备份；实际应用中对可靠性高的部件可不做冗余配置。
  - 软件上具备系统自动检测和切换功能。
  - 错误恢复过程需要短暂时间，所有会有短暂停机。
- 高性能计算(High Performance Computing)集群，简称HPC集群。



目录

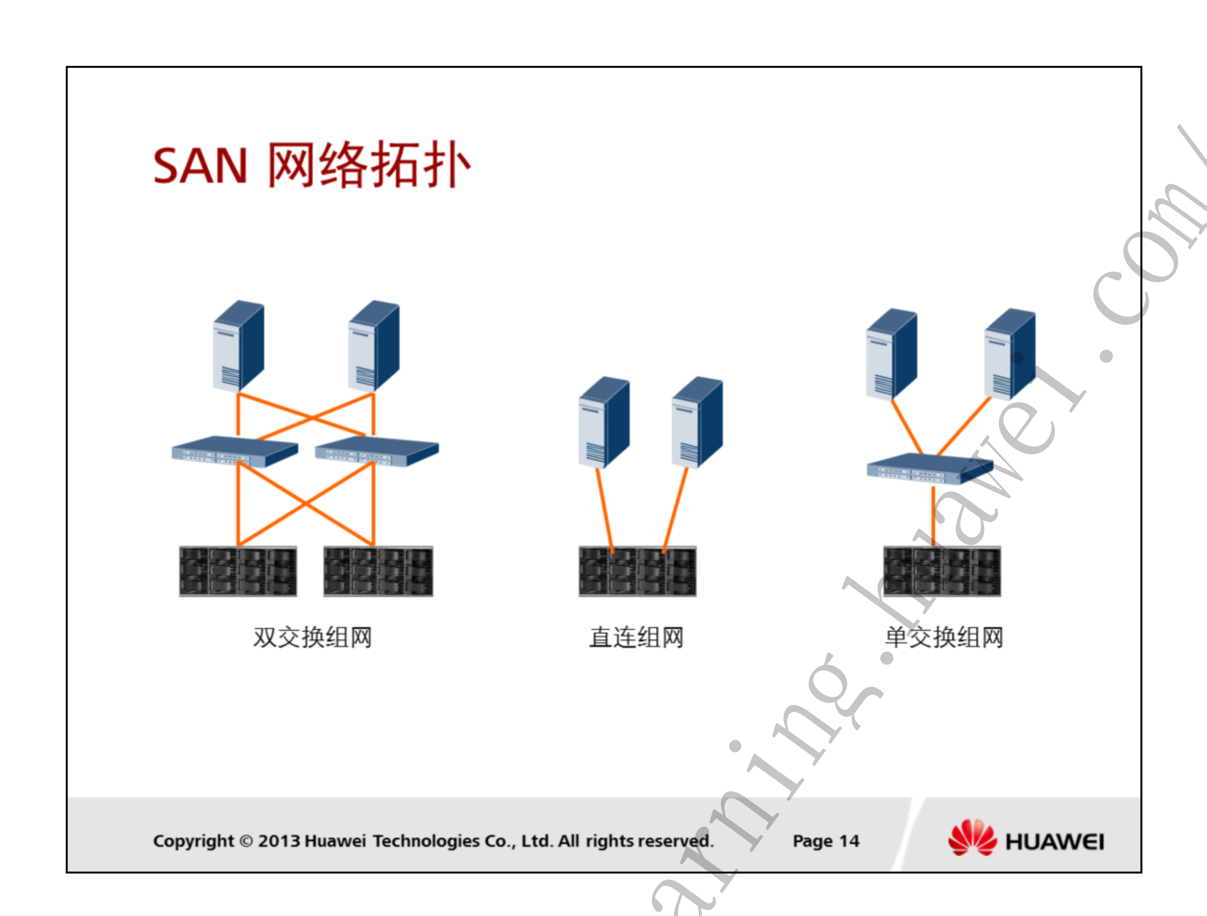


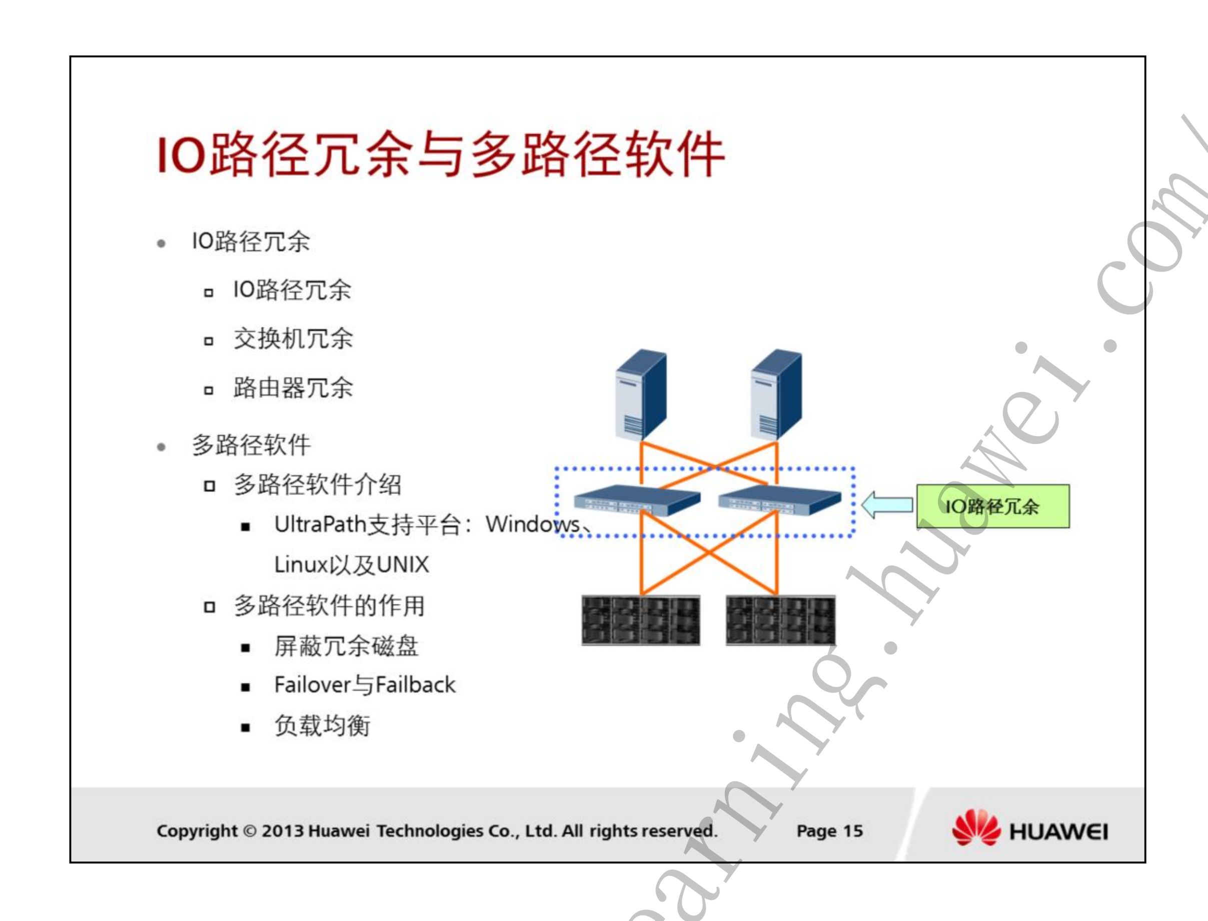
SAN网络架构选择

- IP SAN 与 FC SAN选择

对比项	IP SAN	FC SAN
适用环境	性能要求低；分布式应用，大多数依赖软件实现，稳定性与可靠性随软件复杂度降低	高性能、关键性的数据中心，稳定性好，可靠性高
标准协议	TCP/IP协议，开放性好，完全跨平台文件系统共享	FC协议与专用网络，受光纤距离影响，易形成存储孤岛
安全性	SCSI提供了非常丰富的安全功能：CHAP身份验证能阻止未授权的访问；IPsec能阻止插入、修改和删除操作，并防止被偷听，确保了私密性	依赖专用FC网络，但FC本身没有安全功能
TCO	利用标准TCP/IP以及以太网网络，简单、成本低，技术成熟、维护简单	复杂与昂贵的光纤网络、HBA卡与FC磁盘，需专业维护工程师



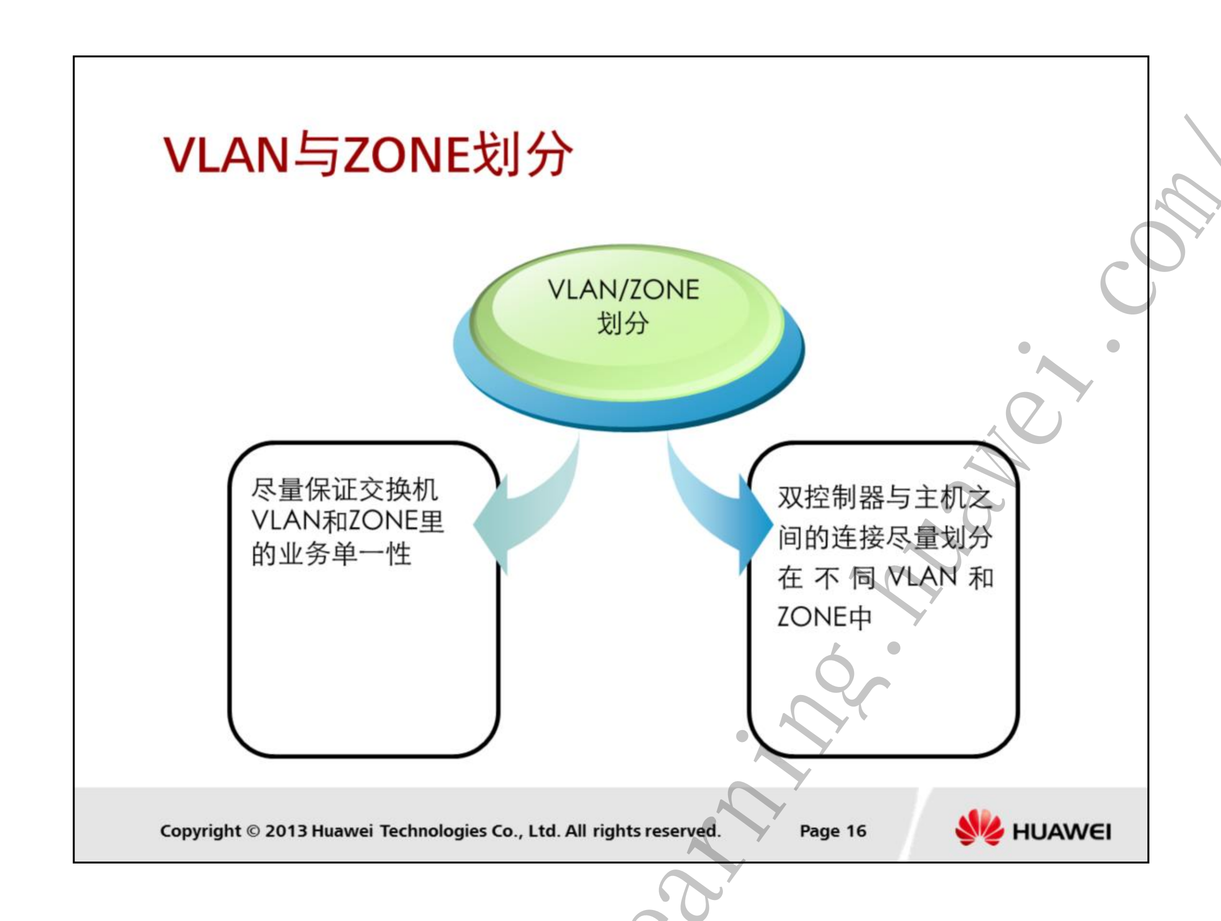

SAN 网络拓扑
双交换组网
直连组网
单交换组网
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 14
HUAWEI



多路径软件支持是为基于应用服务器与 SAN 连接提供高可用性的众多增强功能之一。常用的多路径软件有EMC PowerPath、HP PVLinks、Linux Device-mapper、Microsoft MPIO 。

多路径解决方案使用冗余的物理路径组件（适配器、电缆和交换机）在服务器与存储设备之间创建逻辑路径。如果这些组件中的一个或多个发生故障，导致路径无法使用，多路径逻辑就使用 I/O 的备用路径以使应用程序仍然能够访问其数据。每个网络接口卡（在使用 iSCSI 的情况下）或 HBA 都应通过使用冗余的交换机基础结构连接起来，以便在存储结构组件发生故障时能继续访问存储。

故障转移次数因存储供应商而异，并且可以通过使用 Microsoft iSCSI 软件发起程序驱动程序中的计时器，或修改光纤通道主机总线适配器驱动程序参数设置进行配置。



VLAN，Virtual Local Area Network，即虚拟局域网，是一种将局域网设备从逻辑上划分成一个个网段，从而实现虚拟工作组的数据交换技术 。VLAN划分可防范网络风暴、提高网络安全、降低网络成本等优点。

VLAN常用划分策略，包括基于端口的VLAN划分、基于路由的VLAN划分以及基于MAC地址的VLAN划分，最常用的是基于前两种的划分。

FC SWITCH上的ZONE功能类似于以太网交换机上的VLAN功能，它是将连接在SAN网络中的设备（主机和存储），逻辑上划到为不同的区域内，使得不同区域中的设备相互间不能直接访问，从而实现网络中的设备之间的相互隔离。

- 划分原则：
  - 交换机vlan和zone里的业务单一性，尽量保证在一个vlan或者一个zone中的业务要么是顺序型的，要么是随机型的，确保业务的单一性，提高性能表现。
  - 双控制器与主机之间的连接尽量划分在不同VLAN中，这样可有效预防单点故障。



目录



磁盘选择

- 访问时间=寻道时间+潜伏时间

接口	SATA	SAS	FC
接口类型	串行	串行	串行
拓扑结构	点对点	用扩展器实现的点对点交换式	环路
双工	半双工	双工	双工
磁盘速率 rpm	7200	15000	15000
接口速率 bps	3G、6G	3G、6G	4G
应用	适用于IO负载较轻的应用如文件共享、FTP、多媒体与灾备应用	适于数据库、Email、WEB应用等	适于数据库、Email、WEB应用等

 Page 18



- 平均寻道时间（Average Seek Time）：指硬盘在接收到系统指令后，磁头从开始移动到移动至数据所在的磁道所花费时间的平均值，它一定程度上体现硬盘读取数据的能力，是影响硬盘内部数据传输率的重要参数。平均寻道时间单位为毫秒（ms）。平均寻道时间实际上是由转速、单碟容量等多个因素综合决定的一个参数。一般而言，硬盘的转速越高，其平均寻道时间就越低；单碟容量越大，其平均寻道时间就越低。当单碟片容量增大时，磁头的寻道动作和移动距离减少，从而使平均寻道时间减少，加快硬盘速度。在硬盘上数据是分磁道、分簇存储的，经常的读写操作后，往往数据并不是连续排列在同一磁道上，所以磁头在读取数据时往往需要在磁道之间反复移动，因此平均寻道时间在数据传输中起着十分重要的作用。在读写大量的小文件时，平均寻道时间也起着至关重要的作用。在读写大文件或连续存储的大量数据时，平均寻道时间的优势则得不到体现，此时单碟容量的大小、转速、缓存就是较为重要的因素。
- 平均潜伏时间（Average latency time）：指当磁头移动到数据所在的磁道后，然后等待所要的数据块继续转动到磁头下的时间，盘片转动速度越快，平均潜伏期也就越短，平均潜伏时间单位为毫秒（ms）。
- 平均访问时间 （Average access time）：指磁头找到指定数据的平均时间，通常是平均寻道时间和平均潜伏时间之和（实际上还应该包括一些内部指令操作时间，但这个时间很短，可以忽略不计）。平均访问时间最能够代表硬盘找到某一数据所用的时间，越短的平均访问时间越好。平均访问时间单位为毫秒（ms）。

RAID级别选择

- 常用RAID级别比较：

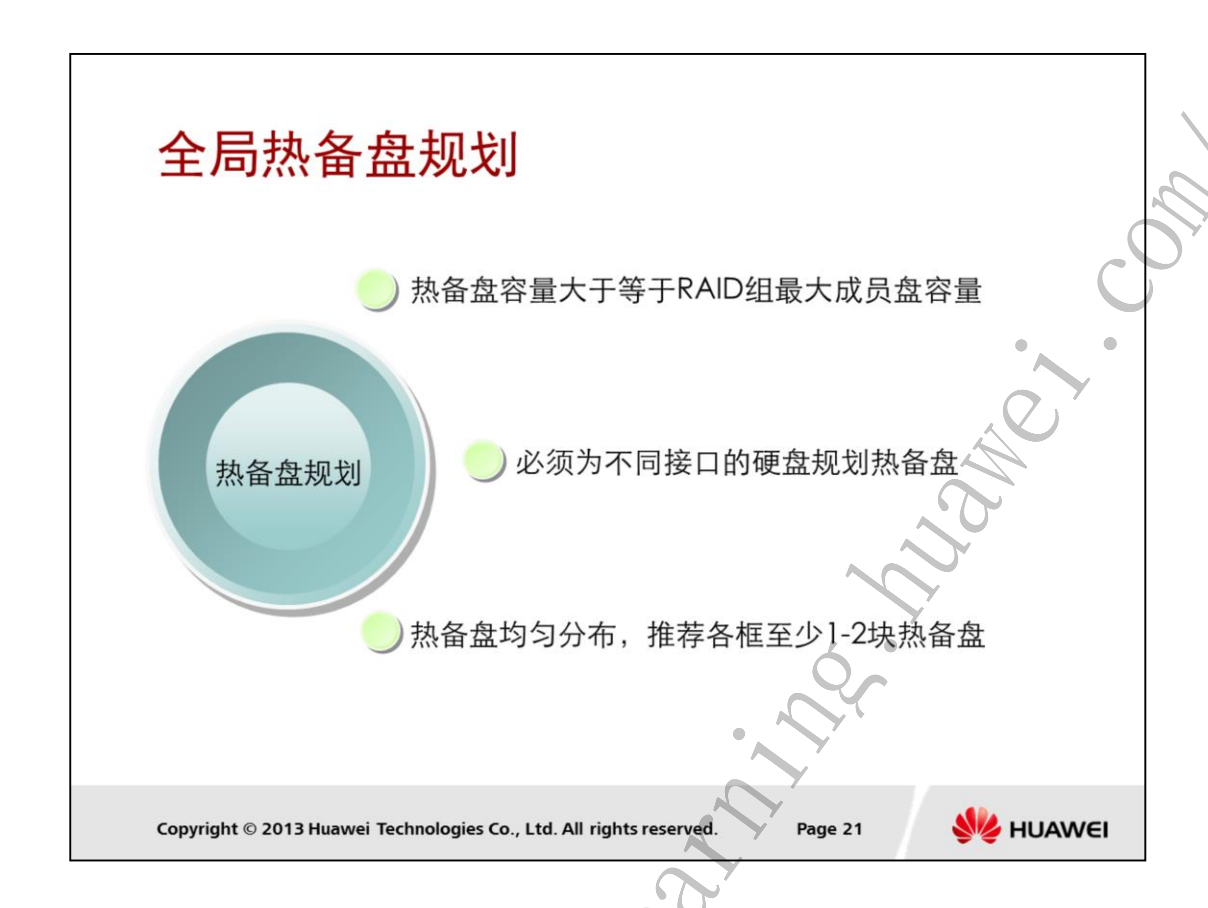
	RAID 0	RAID 1	RAID 10	RAID 5	RAID6
可用容量	总的磁盘容量;	总磁盘容量的1/N; (N：镜像盘个数)	总磁盘容量的1/N; (N：镜像盘个数)	总磁盘容量的(N-1)/N; (N：磁盘数目)	总磁盘容量的(N-2)/N; (N:磁盘数目)
冗余度	无冗余能力，RAID0中一块盘故障，导致数据损坏	100%冗余能力，RAID1中各盘完全镜像	100%冗余能力，RAID10镜像组内镜像	使用XOR校验盘，单盘容错能力	引入两块校验盘，保证在两块盘故障时数据不丢失
典型应用	无故障的迅速读写，对安全性要求不高，如图形工作站等。	随机数据读写，或者要求安全性高，如数据库、服务器存储领域。	随机数据读写，或者要求安全性高，如银行，金融等领域。	随机读写较少、对可靠性要求不是非常高，如视频监控系统等。	对可靠性要求比较高的场景



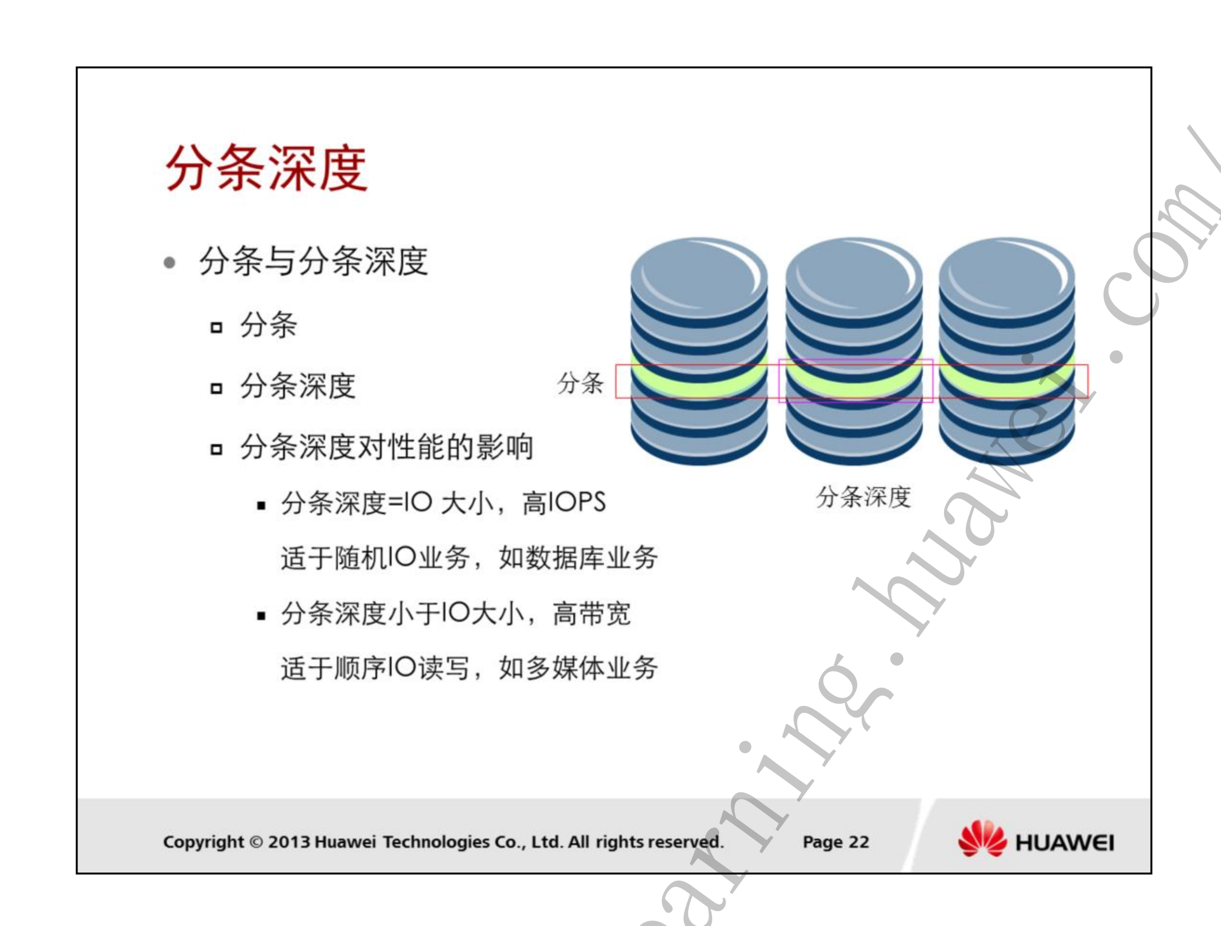
RAID性能

IO模型	RAID10与RAID5对比
顺序读大数据块时	Raid10性能比Raid5更低，但大于Raid5的一半
顺序写	前端压力足够时，Raid5的写带宽会优于Raid10的写带宽
随机读	Raid10的随机读性能与Raid5的随机读性能基本相当
随机写	Raid10的随机写性能优于Raid5
推荐磁盘数	RAID5推荐成员盘数5-9个，RAID10推荐组内镜像盘个数为2个
典型应用推荐	大文件读写操作时，Raid5的性能会明显好于Raid10； 对于随机小数据块读写为主的业务，Raid10是最优的选择

 Page 20 HUAWEI



- 热备盘对RAID0没有数据保护作用。
- 保险箱盘不能作为热备盘使用。



分条是指连续的数据分割成相同大小的数据块，把每段数据分别写入到阵列中不同磁盘上的方法。

分条深度（Stripe Size）：对于一个分条每个成员盘所占的空间我们称为分条深度，现在所支持的分条深度4KB~512KB。如果分条深度太小，很有可能出现一个IO横跨了两个甚至多个分条深度的情况。随着分条深度的增大，一个IO跨盘的几率逐渐减小。因此，随着分条深度的增加，随机读IOPS会逐渐的增加。

当分条深度的大小能够超过数据块的大小达到一定程度时，由于再出现IO数据块跨盘情况的几率已经非常的小了，则数据块的大小已经不再是影响性能的关键因素了，此时随机读IOPS会基本保持一个稳定的水平。

分条深度

- 典型应用IO业务模型
  - Oracle/SQLServer 默认块大小： 8K
  - DB2 默认块大小：4K
  - NTFS/EXT3 默认簇大小：4K

应用	带宽利用率	读/写比例	业务类型	典型IO大小
OLTP、Email、Web、电子商务	低	80% 读，20% 写	随机IO	8K
图像处理、决策支持系统	低	80% 读，20% 写	顺序IO	16K-128K
视频、多媒体业务	高	90% 读，10% 写	顺序IO	大于64K
备份、容灾业务	高	不固定	顺序IO	大于64K



LUN1归属A控制器，LUN2归属B控制器，主机访问LUN1，直接通过控制器A下发访问请求；主机访问LUN2，需要先通过控制器A，然后通过控制器A和控制器B之间的IBS通道，然后再通过控制器B下发访问请求。对LUN1的访问即是本端访问的方式。对LUN2的访问即是对端访问的方式。

对端访问的方式必然要通过板间的IBS通道，因此必然受到IBS通道的限制，所以会出现其读写的性能受到影响的情况。因此，我们在归属LUN的时候，需要注意将该LUN归属给将访问其的主机所连接的控制器上。

双控动态均衡技术：双控制器互为热备，并行处理业务，并动态实现负载均衡，双控之间的Cache镜像采用专用的双通道，双通道之间动态负载均衡

- LUN归属原则：
  - LUN单一归属原则：同一Raid组下的LUN尽量归属同一控制器，特别对于SATA盘，双端访问导致性能下降。
  - 业务均衡分布原则：两个控制器分别归属的LUN个数或者业务压力大小保持基本相当。
  - 业务单一分布原则 ：确保各个控制器的业务类型尽量单一，特别是尽量保证在一个控制器上不同时存在顺序和随机这两种不同类型的业务。

Cache预取策略

不预取

- 读命中，不读磁盘
- 读不命中，向磁盘读该IO
- 适于随机业务读写

固定预取

- 读命中：预取固定大小
- 读不命中：预取大小=不命中长度+固定预取大小
- 适于持续顺序读业务

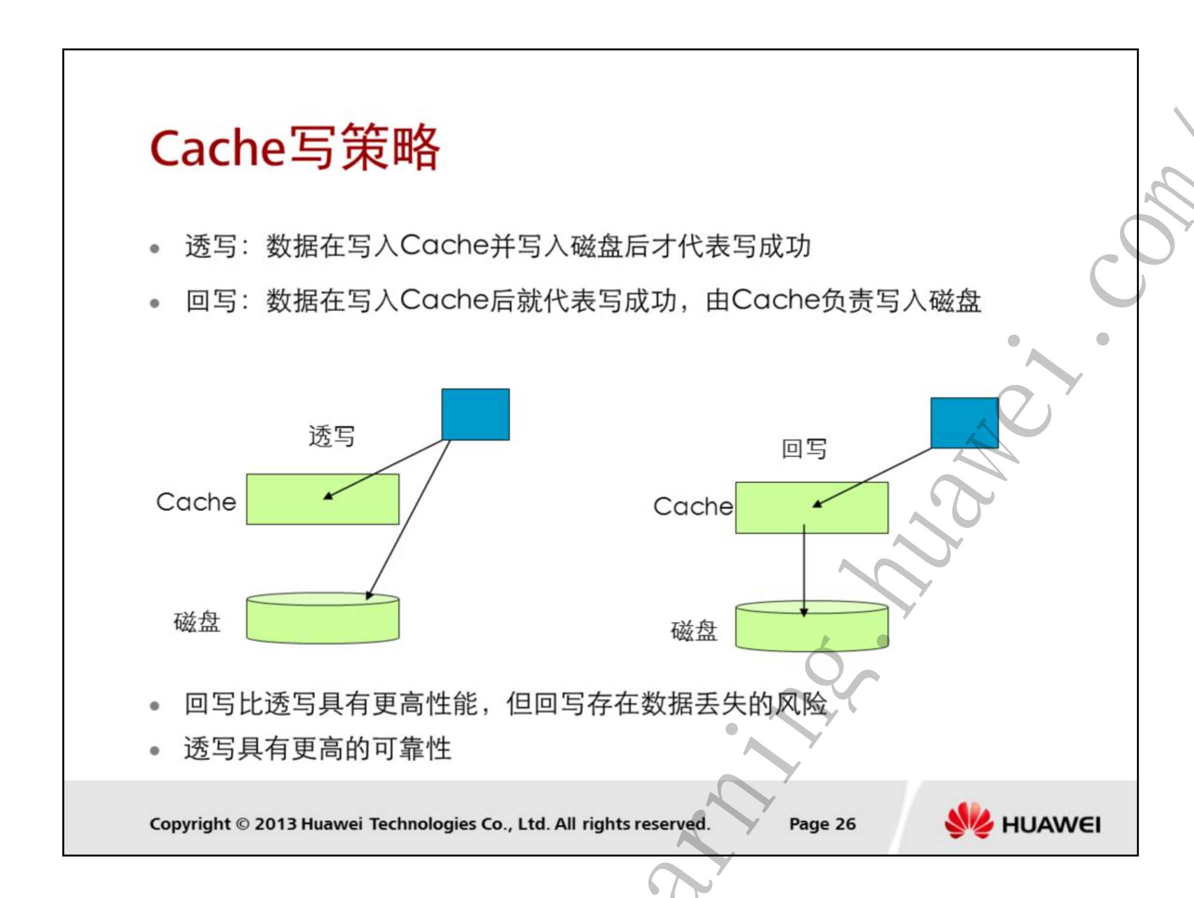
倍数预取

- 读命中：预取长度=预取倍数*读命令长度
- 读不命中：预取长度=不命中长度+预取倍数*读命令长度
- 适于访问的文件或数据库记录从物理地址是离散的，但读取数据又属于连续的

智能预取

- 读命中：预取固定长度
- 读不命中：预取长度=不命中长度+固定预取长度
- 适于持续小数据块读应用





Cache的写策略包括三种：透写，回写镜像和回写不镜像。

- 透写：每一个写IO请求都直接下到磁盘，整体性能的表现受限于磁盘本身所能提供的性能，主要取决于磁盘的转速，寻道时间等关键的参数。
- 回写：每一个写IO到达Cache就代表写入成功；然后通过Cache层面相应的整合等操作，将这些数据对应的装入CHUNK中，再统一下发到磁盘。由于写入Cache的速度会快很多，因此可以得到比透写时更好的性能表现。

回写镜像和回写不镜像这两种写策略的唯一差别就是，当一个写IO从A控制器下发，当到达A控制器的Cache之后，是否需要将该IO再写入到B控制器的Cache作备份。

对于随机的写业务来说：由于业务且流量较小，性能的瓶颈主要在系统资源的分配、后端磁盘的个数、磁盘的转速等因素上，镜像并不是写性能的关键的因素。不管是否采用了镜像的方式，随机写性能几乎没有区别。

如果出现镜像时的随机写性能略微优于不镜像时的随机写性能也是正常的现象。

统一存储系统Hyper功能规划

- 链路规划 → 如果涉及到LUNcopy或者远程复制，需要根据链路带宽、业务类型、产品配置来规划链路
- HostAgent → HostAgent的部署位置、组网规划、策略制定以及与应用软件、操作系统的兼容性规划等
- 资源池 → 资源LUN的大小和数量规划
- 容灾方式 → 如何实现容灾数据的使用


目录


HUAWEI

案例：某中心信息系统SAN存储规划

- SAN业务需求
  - Oracle RAC 业务规格

应用	当前容量	增量	业务特点
财务报表系统	1T	150G/年	数据业务访问量较大，该系统面向全国办公人员，年最大停机时间不得超过10H
财务核算系统	500G	20G/年	每天都有数据操作，在月初与月底时，业务量较大，且不允许中断
资金系统	120G	10G/年	每天都有数据操作
内网门户	50G	5G/年	面向全国办公人员
OA系统	30G	3G/年	政务办公系统，包括公文审批、领导安排、会议管理等

 Page 29



国家某中心信息系统涉及到人、财、物、资金、生产经营等业务系统，其业务数据非常重要，通过深入分析该中心系统建设需求，汇总得出其整体建设目标：整合总部存储资源，提高存储基础架构的可扩展性和性能，满足在三到五年内业务发展，并使整个系统在国内三到五年内保持领先的水平。

案例：某中心信息系统SAN存储规划

- SQL Server集群业务

应用	当前容量	增量	业务特点
审计系统	30G	2G/年	每年预计开展3000个项目的审计，审计工作全年分散进行，年最大停机时间不超过10H
人资系统	10G	1G/年	人员数据稳定不变，每个月变化的主要是保险、培训、招聘等动态信息
惩防系统	1G	100M/年	数据变化量很小，部分附件以文件形式存放在指定目录下
外网门户与BBS	40G	10G/年	外网门户与BBS访问量较大，主要为新闻、评论、图片等



案例：某中心信息系统SAN存储规划

- 建设目标
  - 整合总部存储资源，提高存储基础架构的可扩展性和性能，满足在三到五年内业务发展，并使整个系统在国内三到五年内保持领先的水平。
- 请为其规划SAN方案
  - 要求能够满足SAN性能、可靠性、容量与扩展性需求。



主机规划

- Oracle RAC规划
  - Oracle RAC，推荐采用小型机和RAC集群存储
- SQLServer集群规划
  - SQLServer 集群，推荐采用X86服务器，OS可以采用Windows2003/Windows2008，集群存储推荐NTFS文件系统。
  - SQLServer数据文件与日志文件分布不同的卷/分区上。

 Page 32



- 规划点：性能、可靠性、业务分布、存储容量。

财务报表系统要求高性能，数据库业务主要为随机小数据块业务，衡量SAN性能的主要指标为IOPS。因此在规划SAN前，应先测试一下SAN IOPS峰值为多少，在本次规划中暂不考虑IOPS，仅从总体上定义SAN性能要求。

SAN组网选择

- 性能
  - IP SAN与 FC SAN选择
- 可靠性
  - IO路径冗余与多路径软件
  - UltraPath，支持AIX 与Windows
- 为保证稳定性和性能，推荐使用FC



 Page 33 HUAWEI

容量规划

应用	当前容量	年增量	五年总容量
财务报表系统	1T	150G/年	1750G
财务核算系统	500G	20G/年	600G
资金系统	120G	10G/年	170G
内网门户	50G	5G/年	75G
OA系统	30G	3G/年	45G
Oracle RAC共计：大于2700G			
审计系统	30G	2G/年	40G
人资系统	10G	1G/年	15G
惩防系统	1G	100M/年	1.5G
外网门户与BBS	40G	10G /年	90G
SQLServer 集群共计：大于150G			

总容量：3000GB，即3TB

 Page 34



对于Oracle，由于客户对系统与数据要求较高，因此需开启归档模式，需额外为Oracle提供归档日志空间，可以根据客户需求，保留归档日志多久，本方案定为150G（假定每天产生5G归档日志，保留30天）。

Oracle容量需求约大于2700G，考虑到需为客户预留一定的存储空间（以及CRS、VoteDisk磁盘需1G），规划总容量为3T。

SQLServer容量共计150G。

SAN选择

- 规划点：性能、可靠性、容量、组网
  - Oracle数据库访问量较大，对SAN存储性能要求较高，选择FC SAN，推荐T系列存储。
- 结论：选择S5600T作为SAN存储



硬盘选择

- SAN系统对性能与可靠性要求较高，推荐采用500G SAS磁盘作为后端存储；
- 从SAN总容量需求，推荐选择为标称容量为500G的硬盘；
- 从节省成本考虑，不推荐跨级联硬盘框方式保证SAN容量。



RAID组与热备盘规划

- RAID组
  - Oracle RAC财务报表，业务量较大，性能要求较高，推荐单个规划RAID组

    存储容量需求为1750G，推荐8块SAS盘（实际容量2000G）。
  - 其他业务应用业务性能要求较低，可放置在同一RAID组中。

    存储容量需求为1100G，推荐6块SAS盘（实际容量为1500G）

- 热备盘，推荐2块热备盘



RAID组规划

应用	当前容量	年增量	五年数据总量	RAID组规划
财务报表系统	1T	150G/年	1750G	RAID10_1
财务核算系统	500G	20G/年	600G	RAID10_2
资金系统	120G	10G/年	170G	RAID10_2
内网门户	50G	5G/年	75G	RAID10_2
OA系统	30G	3G/年	45G	RAID10_2
审计系统	30G	2G/年	40G	RAID10_2
人资系统	10G	1G/年	15G	RAID10_2
惩防系统	1G	100M/年	1.5G	RAID10_2
外网门户与BBS	40G	10G /年	90G	RAID10_2
Oracle 归档日志				RAID10_2

RAID组	磁盘容量	磁盘数	镜像组内盘数	总容量
RAID10_1	标称：500G 实际：465G	8	4	标称：2000G 实际：1860G
RAID10_2	标称：500G 实际：465G	6	3	标称：1500G 实际：1395G

 Page 38



所有IO业务都是数据库读写业务，Oracle/SQLServer默认数据块均为8K大小，推荐RAID10配置，镜像组内盘个数采用2个。

整个系统性能需求较高，规划RAID组时，规划成两个RAID组，RAID组名称分别为RAID10_1,RAID10_2，并将业务均匀分配在两个RAID组中（考虑到财务报表系统业务访问非常繁忙，建议单独一个RAID组）。

考虑到保险箱磁盘访问量较大，则不投入应用，可作为后续业务扩展的需求。

LUN规划

RAID组	业务	LUN	LUN容量	LUN配置
RAID10_1	财务报表系统	LUN_FIN_REPORT_1	400G	分条深度：8K 预取策略：不预取 Cache写策略：回写镜像 LUN归属： RAID10_1上LUN全部归属A RAID10_2上LUN全部归属B
		LUN_FIN_REPORT_2	400G	
		LUN_FIN_REPORT_3	400G	
		LUN_FIN_REPORT_4	400G	
		LUN_FIN_REPORT_5	400G	
RAID10_2	财务核算系统	LUN_FIN_CMP_1	750G	
	资金系统	LUN_CAP_1	210G	
	内网门户	LUN_INNET_1	110G	
	OA系统	LUN_OA_1	60G	
	审计系统	LUN_AUDIT_1	50G	
	人资系统	LUN_STU_1	40G	
	惩防系统	LUN_ACK_1	20G	
	外网门户与BBS	LUN_PORTAL_1	120G	
	Oracle 归档日志	LUN_ARCH_1	70G	
		LUN_ARCH_2	70G	





思考题

- 统一存储网络规划的目标有哪些?
- 统一存储网络规划的原则是什么?
- 统一存储网络规划的流程是什么?





总结

- 统一存储系统规划的目标
- 统一存储系统规划的步骤和方法





习题

- 判断题
  1. 在Web应用、Email、数据库应用等小文件频繁读写的业务情况下，SAN性能主要由带宽决定；在视频、测绘等大文件持续读写业务下，SAN性能主要由IOPS决定。（T of F）
- 单选题
  1. 以下哪项不是统一存储规划原则？（　　）

     A.单一需求

     B.经济性

     C.简单设计

     D.封闭性

 Page 42 

- 习题答案：
  - 判断题：1.F
  - 单选题：1.D

Thank you

www.huawei.com

HC120920008 OceanStor 18000 系列存储系统安装及维护



OceanStor 18000系列存储系统安装及维护

www.huawei.com

HUAWEI



目标

- 学完本课程后，您将能够:
  - 了解OceanStor 18000系列存储系统的硬件结构
  - 了解OceanStor 18000硬件安装维护



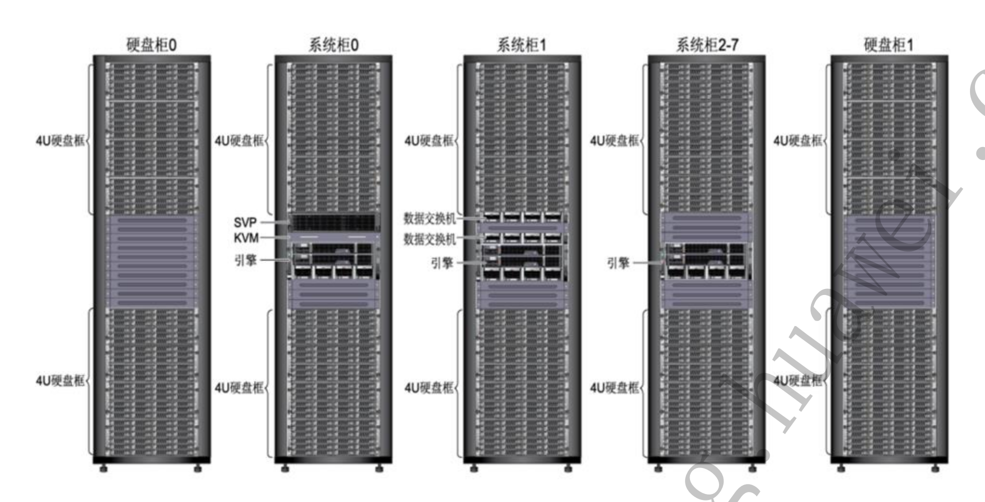


目录




硬件总体介绍
OceanStor 18000系列存储系统外观图
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 3
HUAWEI

硬件总体介绍



OceanStor 18000系列存储系统最高配置前视图（以4U硬盘框为例）

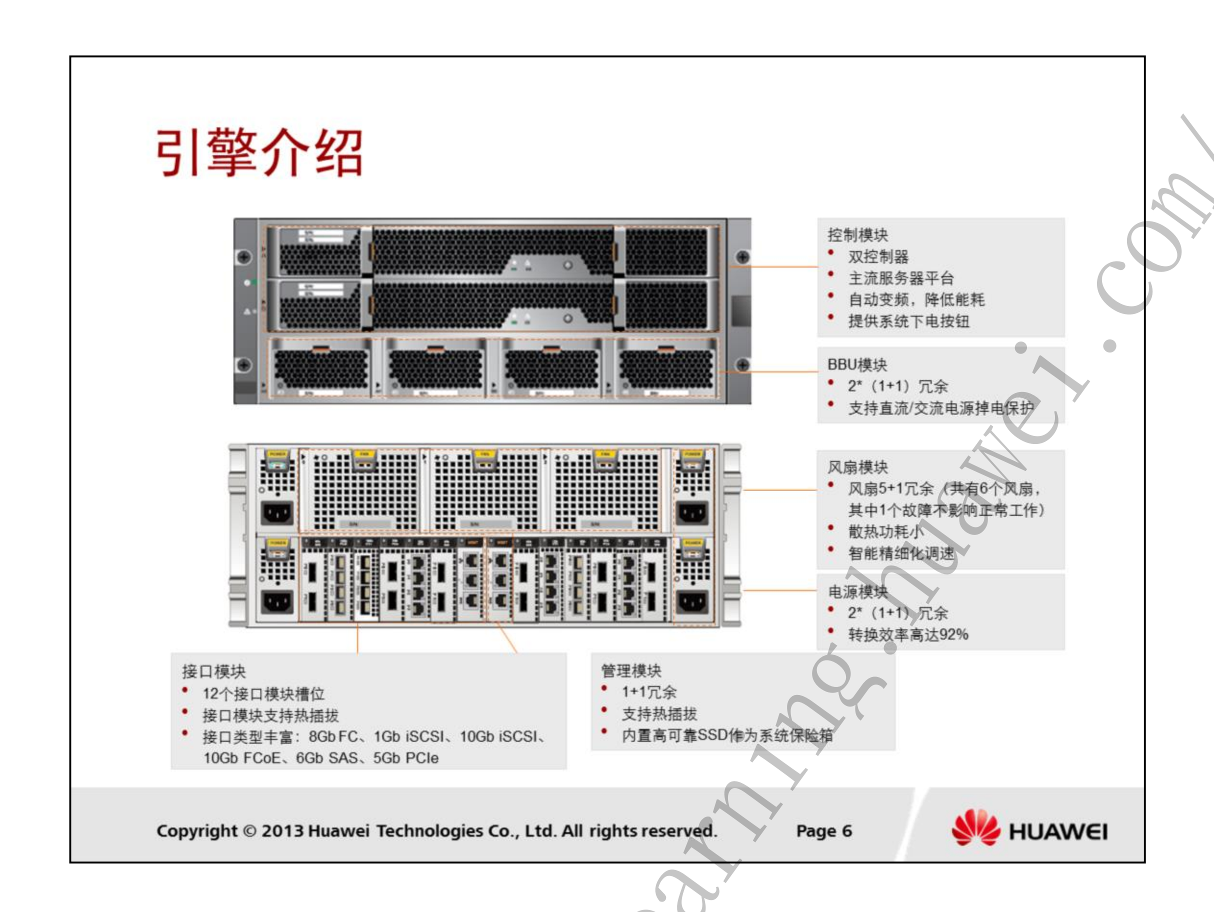


硬件总体介绍

产品类别	产品型号	特性简介
引擎	不涉及	4U独立机头，支持12个接口模块槽位。
硬盘框	4U SAS硬盘框	24*3.5寸硬盘框，支持NL-SAS、SAS、SSD盘。
	2U SAS硬盘框	24*2.5寸硬盘框，支持SAS、SSD盘。
数据交换机	不涉及	实现控制器间控制信息流和业务数据流交换的关键设备。
SVP	不涉及	SVP与KVM配套使用，是存储系统管理、配置、维护等的核心部件。



介绍存储系统所包含的主要硬件。



系统正在走下电流程不允许插拔控制模块或者BBU；

不允许修改系统各模块的firmware；

升级CPLD需要重启系统并等待1分钟左右直到系统自动重启，在此期间不得对控制器断电（拔框/或系统下电）。

电源2*(1+1)冗余：上面两个电源A0和A1为1+1冗余，下面两个电源B0和B1为1+1冗余；

备注：不能同时拔出上部2个电源模块或下部2个电源模块.

风扇5+1冗余：每个风扇模块里面有两个风扇；

保险箱盘：对于盘控分离的控制框，存储系统中第一个硬盘框的前4个硬盘规划为保险箱盘。

保险箱盘用于存放系统重要数据，以及在电源模块故障时保存Cache中的数据。

每一块保险箱盘上用于存放系统重要数据的容量为23 GB，4块保险箱盘共占用92 GB的容量。

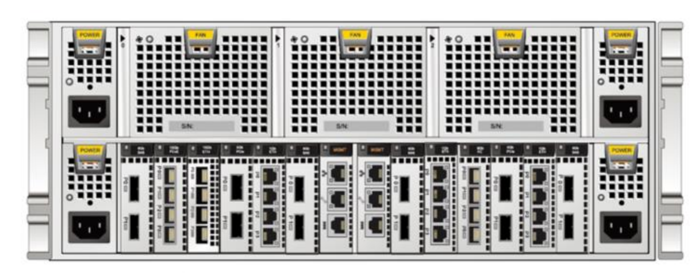
A控2块保险箱盘，B控两块保险箱盘，分别为1+1备份。

引擎介绍

1 3 2 4 5 4

1	管理网口0（接维护终端）
2	管理网口1（接维护终端）
3	串口（接维护终端）
4	mini SAS级联端口（接硬盘框）
5	5Gb PCIe端口（接数据交换机）

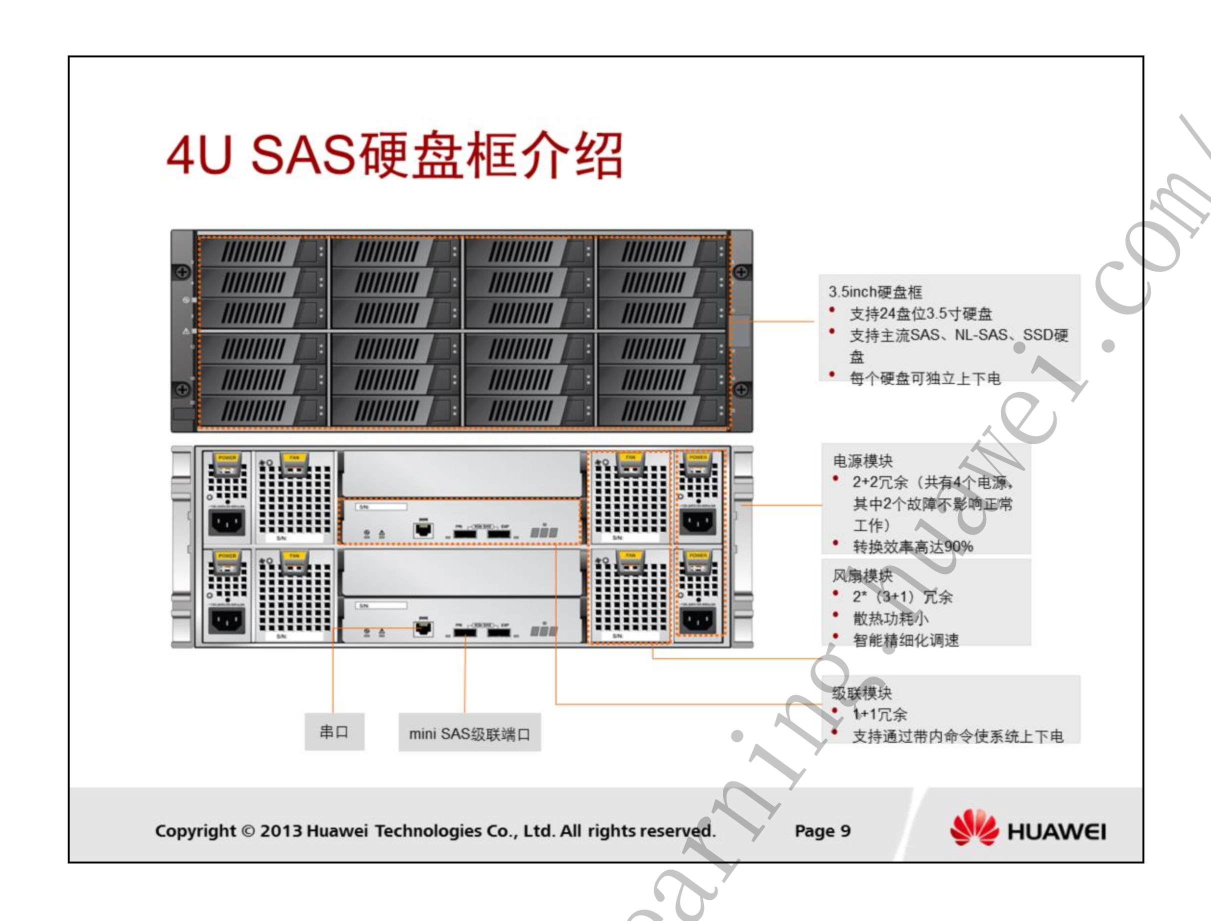
引擎介绍



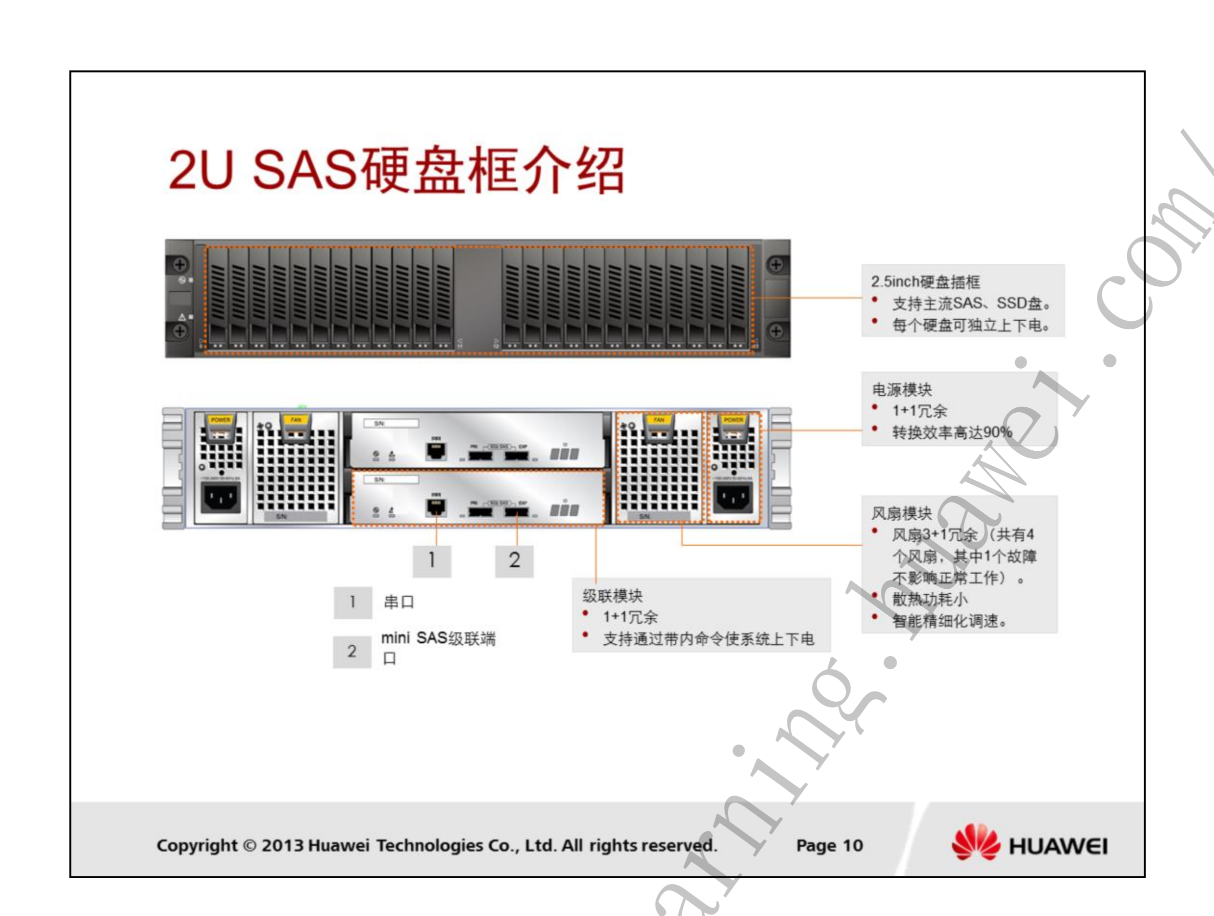
引擎上各个槽位支持的接口模块如下：

- A0～A5由控制器0管理，B0～B5由控制器1管理。
- A0和B0插槽配置6Gbit/s SAS级联模块（必选），用于连接柜中的硬盘框。
- A5和B5插槽为保留槽位，仅在6Gbit/s SAS级联端口不够用的时候，可配置6Gbit/s SAS级联模块。
- A3和B3插槽建议作为预留槽位，在配置有多个系统柜的情况下，可配置5Gb PCIe 接口模块用于连接系统柜1中的数据交换机。
- A1和B1、A2和B2、A4和B4这三对插槽为可选，客户可根据组网情况选择相应的I/O接口模块，包括1Gbit/s iSCSI接口模块、8Gbit/s FC接口模块、10Gbit/s iSCSI接口模块和10Gbit/s FCoE接口模块。



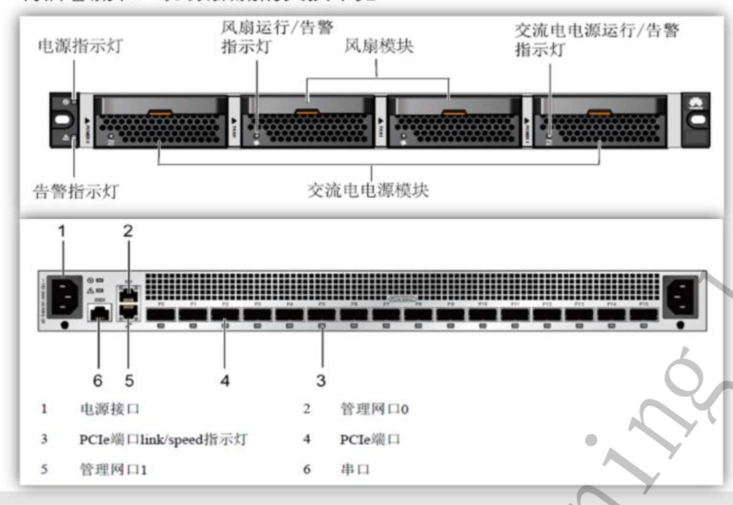


- 级联口必须按上一级的下行口(EXP)接下一级的上行口(PRI)的连接方式进行；
- 升级CPLD时需要停止磁盘框业务；
- 4U标配硬盘框必须插双电源工作；
- 风扇2*(3+1)冗余：每个风扇模块有两个风扇；


2U SAS硬盘框介绍
2.5inch硬盘插框
支持主流SAS、SSD盘。
每个硬盘可独立上下电。
电源模块
1+1冗余
转换效率高达90%
风扇模块
风扇3+1冗余（共有4个风扇，其中1个故障不影响正常工作）。
散热功耗小
智能精细化调速。
级联模块
1+1冗余
支持通过带内命令使系统上下电
1
2
1 串口
2 mini SAS级联端口
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 10
HUAWEI

数据交换机介绍

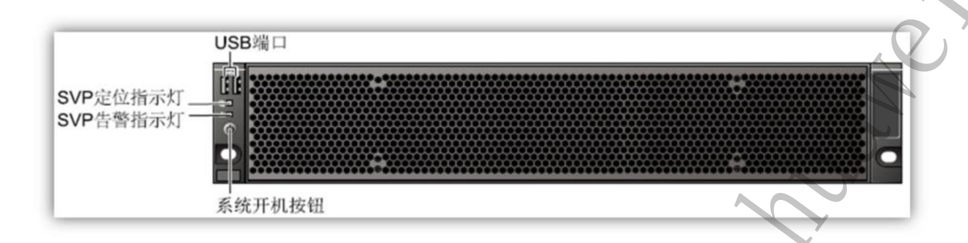
OceanStor 18000系列存储系统中，数据交换机由SVP统一监控和管理。仅在系统柜1上配置了两个数据交换机，两者采用Active-Active的工作模式，提高了整个控制信息流和业务数据流的交换带宽。



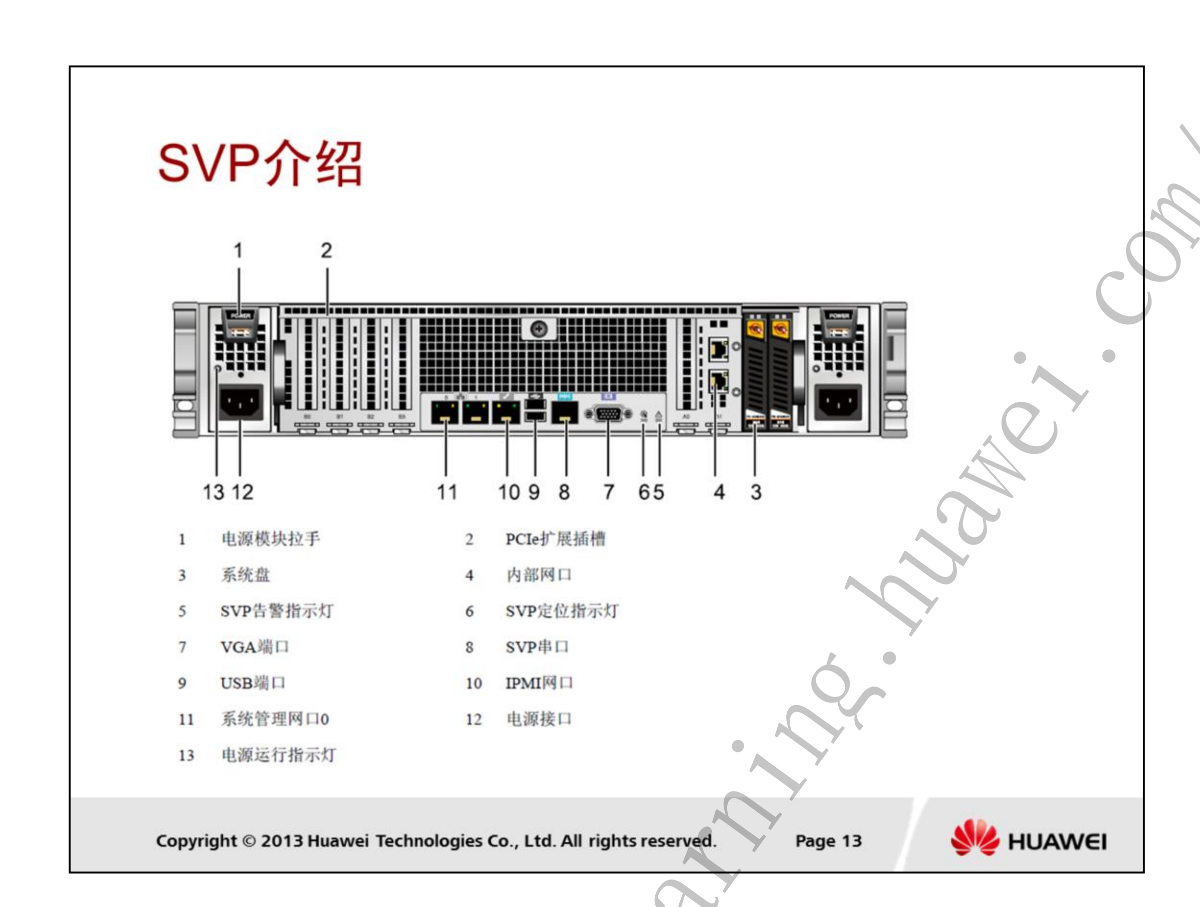
HUAWEI

SVP介绍

SVP与KVM配套使用，是OceanStor 18000系列存储系统管理、配置、维护等的核心部件。其上安装了OceanStor 18000系列存储系统所需的维护、管理等工具，可以在本地或远程轻松完成全套的监控、管理、配置、鉴权等一系列工作。



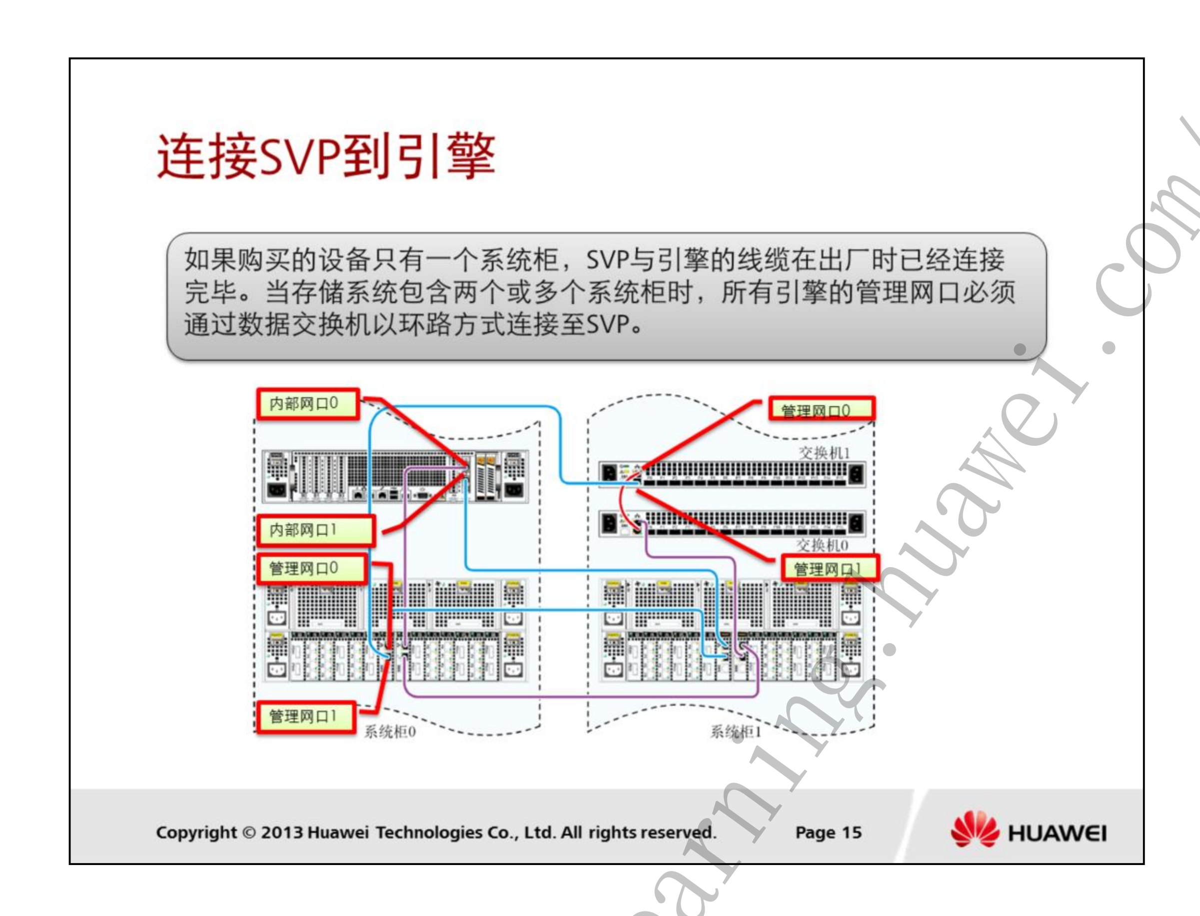



SVP介绍
1
2
13 12
11
10 9
8
7
6 5
4
3
1 电源模块拉手
2 PCIe扩展插槽
3 系统盘
4 内部网口
5 SVP告警指示灯
6 SVP定位指示灯
7 VGA端口
8 SVP串口
9 USB端口
10 IPMI网口
11 系统管理网口0
12 电源接口
13 电源运行指示灯
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 13
HUAWEI



目录





连接步骤：

- 连接数据交换机0的网络网口1到数据交换机2的管理网口0。
- 依次连接SVP的内部网口0到引擎0的A控制器管理网口0、引擎0的A控制器管理网口1到引擎1的A控制器管理网口0、引擎1的A控制器管理网口1到引擎2的A控制器管理网口0……引擎N的A控制器管理网口1到数据交换机0的管理网口0。（N代表引擎的数量）
- 依次连接SVP的内部网口1到引擎N的B控制器管理网口0、引擎N的B控制器管理网口1到引擎N-1的B控制器管理网口0、引擎N-1的B控制器管理网口1到引擎N-2的B控制器管理网口0……引擎0的B控制器管理网口1到数据交换机1的管理网口1。（N代表引擎的数量）

注意：所有连线必须按照以上的规则和顺序进行连接。

连接数据交换机到引擎

如果购买的设备只有一个系统柜，安装时不需连接。当存储系统包含两个或多个系统柜时，必须将各系统柜中的引擎通过PCIe端口连接至系统柜1中的数据交换机。

交换机1
交换机0
系统柜0
系统柜1

 Page 16 HUAWEI

连接步骤：

- 连接数据交换机0到引擎的AOC线缆

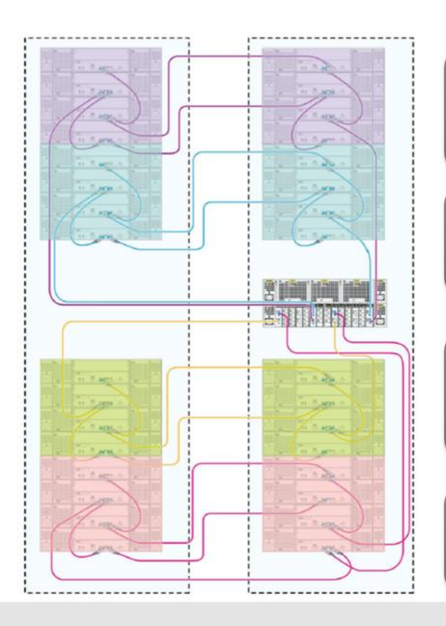
  数据交换机0的P0~P15端口依次连接到的引擎PCIE端口为：引擎0 A3槽位的P0、引擎0 B3槽位的P0、引擎1 A3槽位的P0、引擎1 B3槽位的P0的P1、引擎2 A3槽位的的P0、引擎2 B3槽位的P0、引擎3A3槽位的P0、引擎3 B3槽位的P0、引擎4 A3槽位的P0、引擎4 B3槽位的P0、引擎5 A3槽位的P0、引擎5 B3槽位的P0、引擎6 A3槽位的P0、引擎6 B3槽位的P0、引擎7 A3槽位的P0、引擎7 B3槽位的P0。

- 连接数据交换机1到引擎的AOC线缆

  数据交换机1的P0~P15端口依次连接到的引擎PCIE端口为：引擎0 A3槽位的P1、引擎0 B3槽位的P1、引擎1 A3槽位的P1、引擎1 B3槽位的P1、引擎2 A3槽位的的P1、引擎2 B3槽位的P1、引擎3 A3槽位的P1、引擎3 B3槽位的P1、引擎4 A3槽位的P1、引擎4 B3槽位的P1、引擎5 A3槽位的P1、引擎5 B3槽位的P1、引擎6 A3槽位的P1、引擎6 B3槽位的P1、引擎7 A3槽位的P1、引擎7 B3槽位的P1。

注意：所有连线必须按照以上的规则和顺序进行连接。

连接SAS线缆





如果同时配置有硬盘柜和系统柜，需要预先拆除系统柜与硬盘柜相邻一面的两扇机柜侧门（系统柜 与硬盘柜分别拆除一扇，使SAS线缆能够从机柜之间穿过过）。

控制框上的级联口必须与硬盘框上的PRI级联口连接。可以根据线缆标签找到对应的接口连接即可。

硬盘框上的EXP级联口必须与其它硬盘框的PRI级联口连接。 每个硬盘框级联环路最多能级联4个硬盘框。如果是链路的最末端磁盘框，EXP是悬空的。

同一级联环路中的硬盘框应形成两个相互独立、互为冗余的链路，以达到最佳的组网可靠性。

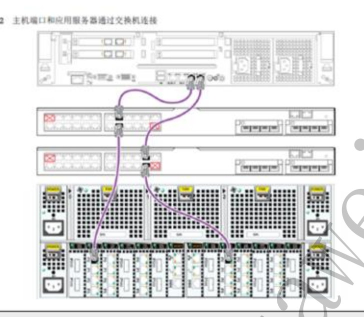
HUAWEI

连接引擎到应用服务器

场景1 主机端口和应用服务器直连



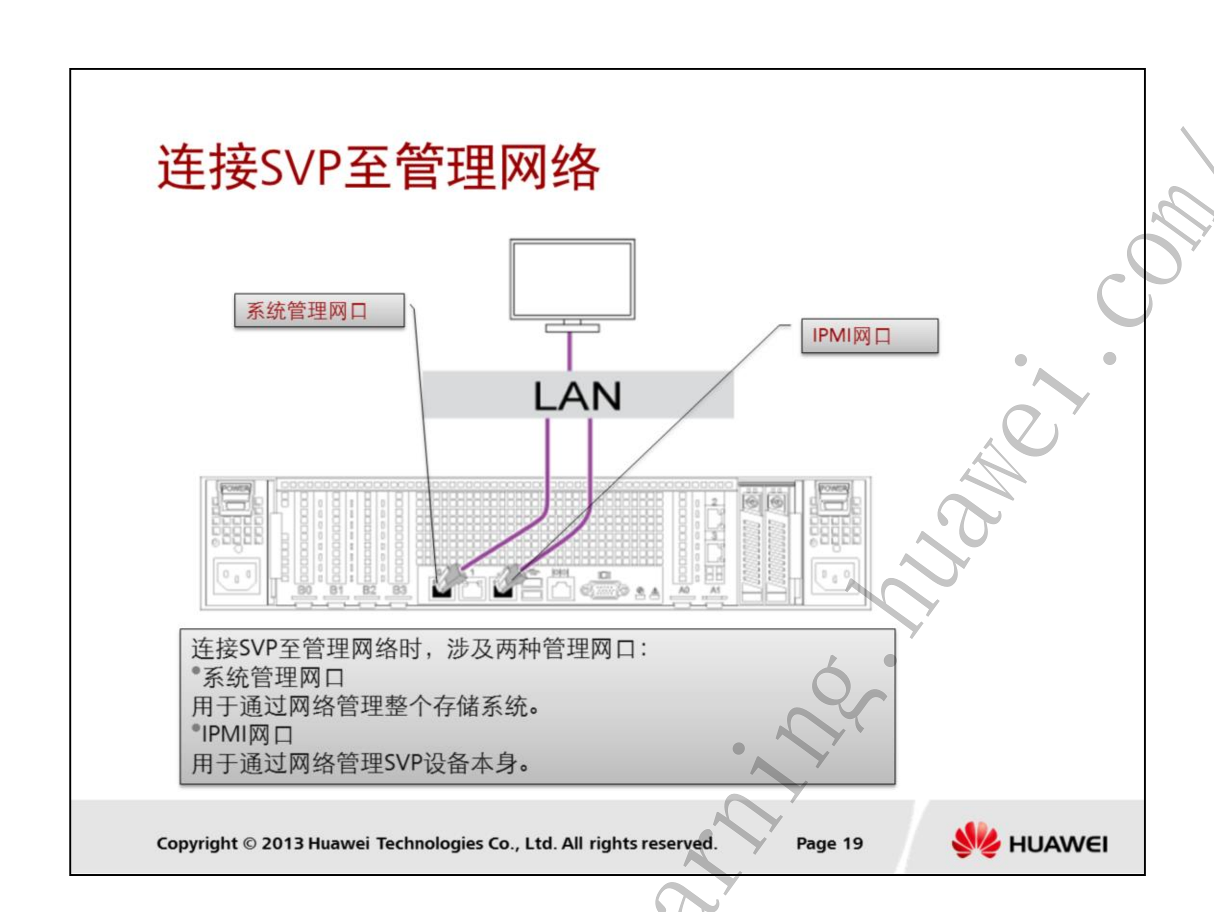
场景2 主机端口和应用服务器通过交换机连接



- 存储系统支持10Gbit/s TOE、8Gbit/s FC、10Gbit/s FCoE以及1Gbit/s接口模块，同一个引擎的两个控制器相同槽位接口模块相同。
- 0号槽位的接口模块用于安装SAS接口模块，5号槽位建议预留， 在0号槽位的6Gb SAS级联端口不够用的时候，可配置6Gb SAS级联模块用于连接硬盘框。
- 3号槽位建议作为预留槽位，在配置有多个系统柜的情况下，可配置5Gb PCIe接口模块用于连接系统柜1中的数据交换机。



 Page 18 HUAWEI


连接SVP至管理网络
系统管理网口
IPMI网口
LAN
连接SVP至管理网络时，涉及两种管理网口：
•系统管理网口
用于通过网络管理整个存储系统。
•IPMI网口
用于通过网络管理SVP设备本身。
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 19
HUAWEI



目录



工程勘测准备

- 了解局点的设备配置情况，如系统柜的数量、框的数量等
- 工具准备：直尺、水平仪、卷尺、三角板、笔
- 打印工程勘测的数据表，以便在勘测过程中填写和记录勘测的数据
- 与局点确认工程勘测时间以及卸货时间（部分城市存在大货车限行的情况）



卸货场地勘查

机柜包装后高度、宽度和深度分别为：2300mm、1100mm和1500mm，带包装总重量为834kg。在将存储系统从运输工具上卸载时，需要注意以下事项：

- 卸载场地需满足存储系统的短时间存放和方便转运的要求，在存放过程中请不要拆卸外包装箱。
- 为方便将拆卸机柜的包装箱，包装箱前后方向至少需要6.5m空间，左右方向至少需要2.0m空间。





注意：

如果局点的环境不满足设备的卸货要求，请及时通知客户并提前确定好卸货地点。



机柜搬运通道要求

机柜的尺寸为1995mm（高）x 600mm（宽）x 1135mm（深），建议的搬运通道尺寸如下表所示。

检查项目	宽度	高度	深度	承重
电梯	≥800mm	≥2100mm	≥1300mm	≥1200kg
通道门	≥800mm	≥2100mm	N/A	N/A

注意：

- 搬运通道不能有坑、槛等阻塞机柜推行的障碍，行经地面要求平整，不能有纸、地毯等。
- 如果搬运过程中需要经过台阶，所有台阶处必须搭建斜台，且倾斜角度不超过10度。
- 机柜拆除包装后，严禁使用叉车运载和升举过台阶。
- 如果搬运过程中需要经过斜坡，请确保斜坡的坡度低于10度。



机房散热要求

因空调需求和机房面积以及设备散热相关，请参考相关的工程设计规范书，确定机房的制冷是否满足存储设备的散热要求。

存储系统的散热量

机柜名称	散热量
系统柜	20460BTU/H
硬盘柜	15345BTU/H

存储系统的散热风量

机柜名称	散热风量
系统柜	660CFM
硬盘柜	480CFM

注意：
如果机房采用下送风，在机柜部署时需要确保每个机柜前有1块通风地板，且地板的通风率要求≥50%，以确保机柜可以获得足够的风量来冷却设备。



温度和湿度要求

存储系统对温度和湿度的要求

参数	条件	要求
温度	工作温度	● 海拔低于1800m时，5℃～35℃ ● 海拔为1800m～3000m时，5℃～30℃
	存放温度（带外包装的运输和存放）	- 20℃～ + 60℃
	工作环境温度变化率	10℃/H
	非工作环境温度	0℃～50℃
	非工作环境温度变化率	10℃/H
湿度	工作湿度	20% RH～80% RH（不冷凝）
	存放湿度（带外包装的运输和存放）	5% RH～95% RH
	非工作环境湿度	10% RH～90% RH
	最大湿度变化率	25%/H

注意：
为确保存储系统长期、可靠的运行，请确保机房的温度和湿度满足设备的运行要求。



机房承重要求

存储系统重量参数

设备名称	重量
系统柜	● 满配2.5寸硬盘：658kg ● 满配3.5寸硬盘：654kg
硬盘柜	570kg
空机柜	178kg
引擎	44kg
2U SAS硬盘框	● 20kg（配置2.5寸SAS硬盘） ● 14.9kg（不包含硬盘）
4U SAS硬盘框	● 44kg（配置3.5寸SAS硬盘） ● 25.2kg（不包含硬盘）

按机柜并柜场景，前后走道间距1m，机柜前后各有一个人同时维护（每个体重100KG）计算，承重要求为：$(653+100*2)/[0.6*(1.135+1/2*2)]=666KG/m^2$

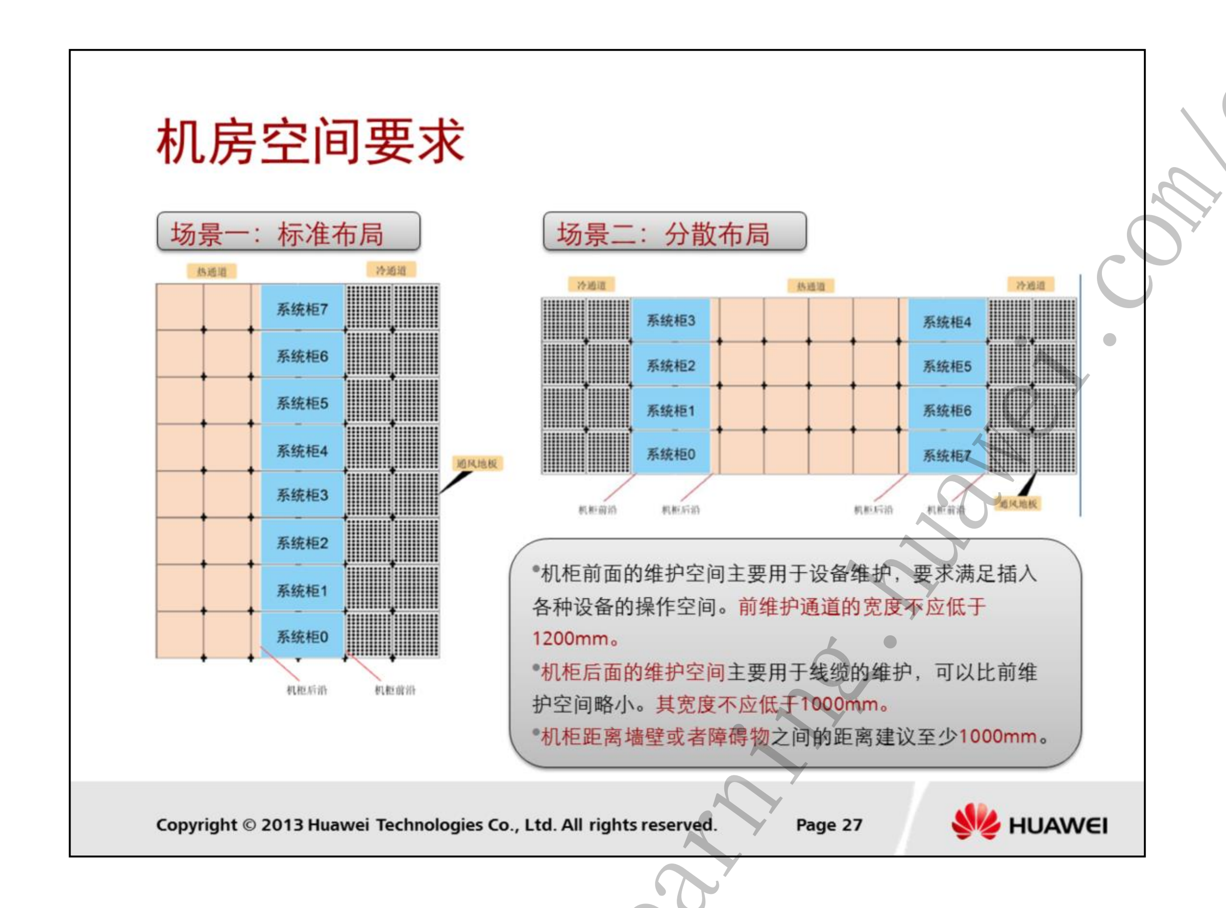
注意：

具体数据可按照机房实际情况计算，计算承重时要求客户考虑安全系数。

 Page 26



存储系统整柜形态交付，重量以整柜重量为最小单位，机柜满配的最大重量为653kg。机柜的尺寸为：1995mm（高） x 600mm（宽） x 1135mm（深）。



- 机柜的尺寸为1995mm（高） x 600mm（宽） x 1135mm（深），机柜采用前门为单开门，后门为双开门，且前门开门方向为从左向右。可以根据机房的实际环境按照标准布局或分散布局进行机柜的安装。请根据机柜的摆放方式以及机柜的数量计算机柜占用的空间，如果后续有扩容需求，建议提前规划并预留。
- 每个系统柜上配置有一台引擎，由系统柜1中的数据交换机汇聚连接所有引擎的AOC线缆，布线和机柜布局应当围绕着系统柜1来展开。

机房供电要求



- 对于北美及110V电源区域，要求需提供满足NEMA标准的三相四线50A（50A 250V 3 PHASE 3-POLE 4-WIRE）电源连接器作为机柜供电外部接口，如不能提供，则需在备注中说明。
- 确定电源线走线路由时，要注意电源线不能从机柜要扩容方向一侧上线，否则下次扩容时新增机柜将压住原来的电源支线，造成无法安装的后果。
- 对于扩容局勘测，需要注意在扩容并柜时是否有电源线从扩容方向走线的情况，如果存在这种情况，最好补发电源支线进行整改。否则到了工程现场将无法安装。
- 如果电压稳定性不能满足要求，应采用调压或稳压设备使电压满足波动范围要求。
- 建议每个机柜的两个PDU的输入电源来自两个独立的配电屏，并且具备独立的UPS（Uninterrupted Power Supply）系统。
- 如果局点的配电环境不满足OceanStor 18000系列存储系统的标准用电要求（输入路数、电源制式、电压情况、输入电流、输入电源线是否含接地等），则需特别备注。

线缆长度及走线要求

- **网线**（用于连接SVP至管理网络，SVP与引擎）
  如果10m不能满足跨接要求，需特别说明数量和长度，指导公司发货。
- **AOC线缆**（用于连接引擎与数据交换机）
  - AOC线缆目前最长尺寸为15m，在进行机柜布局时一定要注意系统柜1与最远的机柜之间的直线距离不能超过10m。
- **SAS线缆**（用于连接引擎与硬盘框）
  - 系统柜内部的SAS线缆出厂时候已经连接完毕。
  - 当配置了硬盘柜时，需要跨柜连接SAS线缆。
  - 跨柜级联的mini SAS线缆默认配置长度为3m。
- **光纤**（用于连接引擎与应用服务器）
  - 光纤主要有10m和3m两种规格。
  - 如果10m长度的光纤无法满足使用要求，需特别说明数量和长度，指导公司发货。

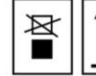
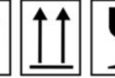
Page 29

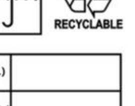


- 线缆走线方式可以选择上走线和下走线，走线方式由客户确定。
- 如果采用上走线，需要现场确认走线架高度是否满足机柜高度要求。必须满足走线架下表面到机房地表面（或防静电地板上表面）的距离大于3米。
- 如果采用下走线，机房必须为防静电地板地面，机房净高度指天花板到防静电地板上表面的距离，要求不小于2.6米。同时需确认机柜摆放位置的下方是否有电缆、孔洞等障碍以利于设备安装。有活动地板的机房应安装底座。
- 当防静电地板不满足承重要求时，需要安置防静电地板支架。防静电地板支持的高度范围为250-400mm，如防静电地板的高度不在此范围内，则需要定制支架，优选推动客户自行设计安装此支架，以保持与原机房环境的协调性和一致性。

包装示意图及标识

储运标识、重心位置、重货&防踩踏标识、开箱指引等。

警告 | 重型货物,倾覆危险

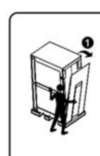
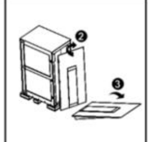
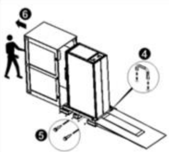
内尺寸(I.D.)	
外尺寸(O.D.)	



重型货物 Top-heavy cargo

严禁堆叠，严禁踩踏，No stacking & No step on

拆箱指导


防冲击监示标签、防翻倒监示标签



尺寸:1500*1100*2300mm 毛重:854Kg




包装方案示意图（以实物为准）

Page 30 HUAWEI

异常场景：拆机柜搬运.

适用场景:

机柜途经大于10度斜坡、台阶、低于2米高的门洞。

解决步骤及方案:

1. 拆卸设备和门:
   ① 拆下柜内设备(有两种方式)
   方式一（推荐）：引擎、硬盘框带硬盘一起拆卸，做好顺序标记；其他线缆和部件保留，剩下250kg左右。
   方式二（不推荐）：在有包装条件的前提下，拆卸机柜正面的FRU部件（包括硬盘、控制板、BBU等模块），机柜后面的部件不做拆卸。2.5寸盘0.2kg，3.5寸盘0.7kg，2个控制板+4个BBU：10kg 。
   ② 拆卸机柜柜门：拆除两个侧门，尽量保留前后门。如右侧图1所示。
2. 搬运：机柜侧倾，把机柜搬运到指定位置。搬运人力至少6个人。
3. 恢复：把拆除的框安装回对应的机柜、对应的位置。线缆按照标签指示插回原来对应的接口。

注意事项

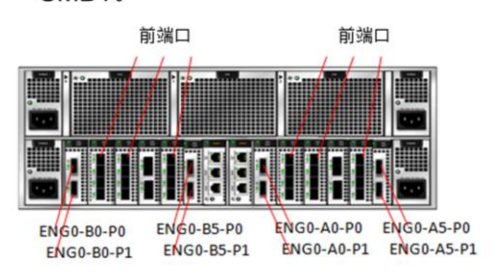
Page 31

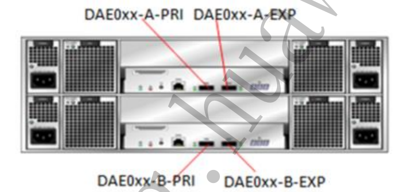
HUAWEI

异常场景：第3步恢复的补充说明1

A. 如何确定硬盘框与机柜的对应关系？

在硬盘框级联模块上，粘贴有级联模块标签，表示此级联模块的端口信息，如DAE011-A，表示此硬盘框的A级联盒。其中的“0”与机柜编号的SMB0的“0”对应，如是DAE100-A,则表示此硬盘框对应的机 柜编号为SMB1。







异常场景：第3步恢复的补充说明2

B. 如何确定设备在机柜中的位置？

在机柜的前后侧壁上，粘贴有维护地图，上面标记各设备在机柜上对应的安装位置。

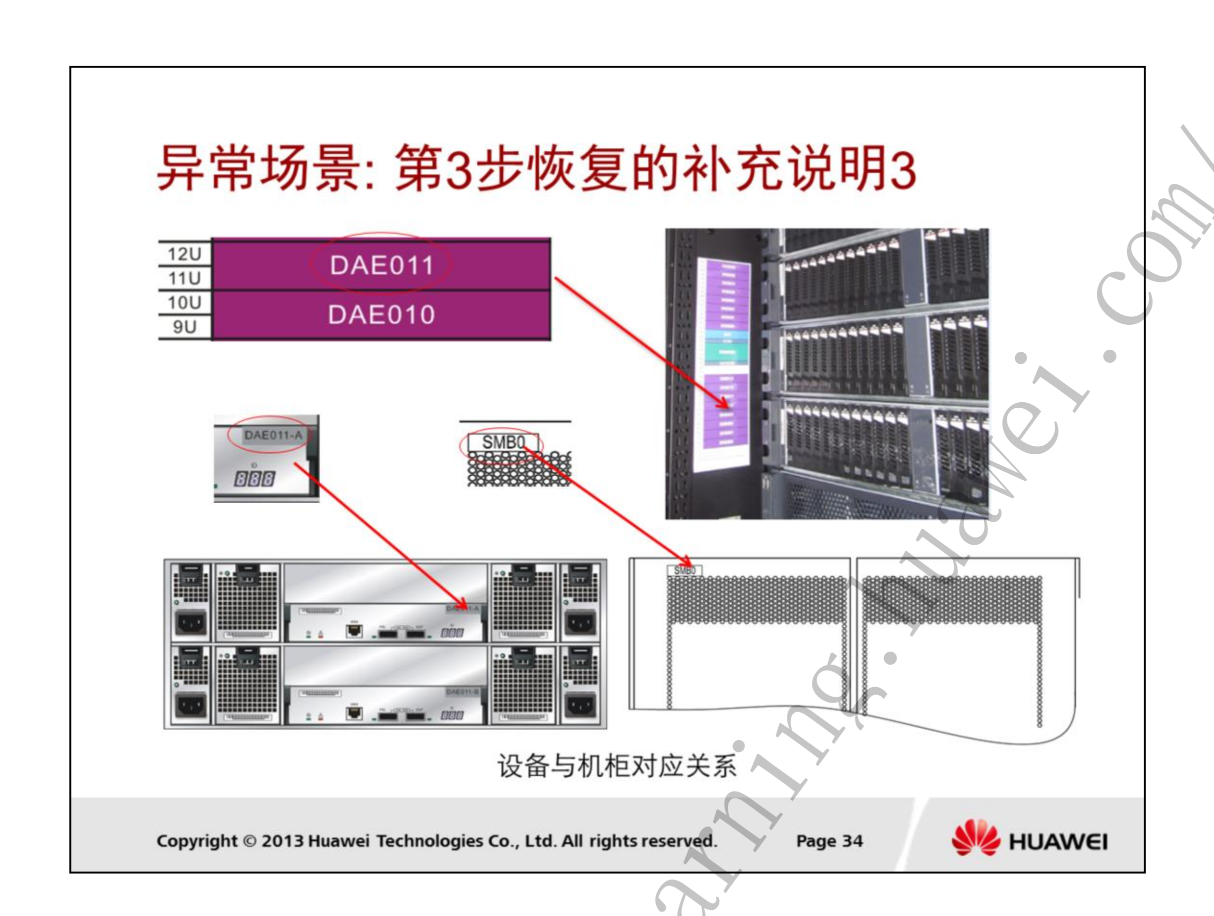
C. 如何确定线缆与设备端口的对应关系？

线缆（电源线缆和信号线缆）上粘贴有标签，表示线缆与设备端口的对应关系。设备端口信息在各设备上均有标示，可以据此找到对应关系。

注意：

1. 对于扩容场景，机柜和设备上可能没有对应标签或标签信息不全，在拆卸前请手动标记好设备与机柜的对应信息。
2. 设备拆除时请整框下柜，切勿拆卸框内的模块及硬盘组件。因KVM安装比较复杂，建议在搬运时不拆除。
3. 为便于搬运，可将侧门、后门拆除。SMB0 的前门不可拆卸，其它柜前门可拆。搬运过程中，前门及前门装饰条不能受力。




异常场景: 第3步恢复的补充说明3
12U
11U
10U
9U
DAE011
DAE010
DAE011-A
SMB0
设备与机柜对应关系
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 34
HUAWEI



目录



OceanStor Toolkit简介

OceanStor Toolkit是由华为技术有限公司开发的工具，通过该工具可以帮助技术服务工程师、运维工程师对设备进行部署、维护和升级。

- **部署功能**
- **维护功能**
- **升级功能**

 Page 36 

- 部署功能

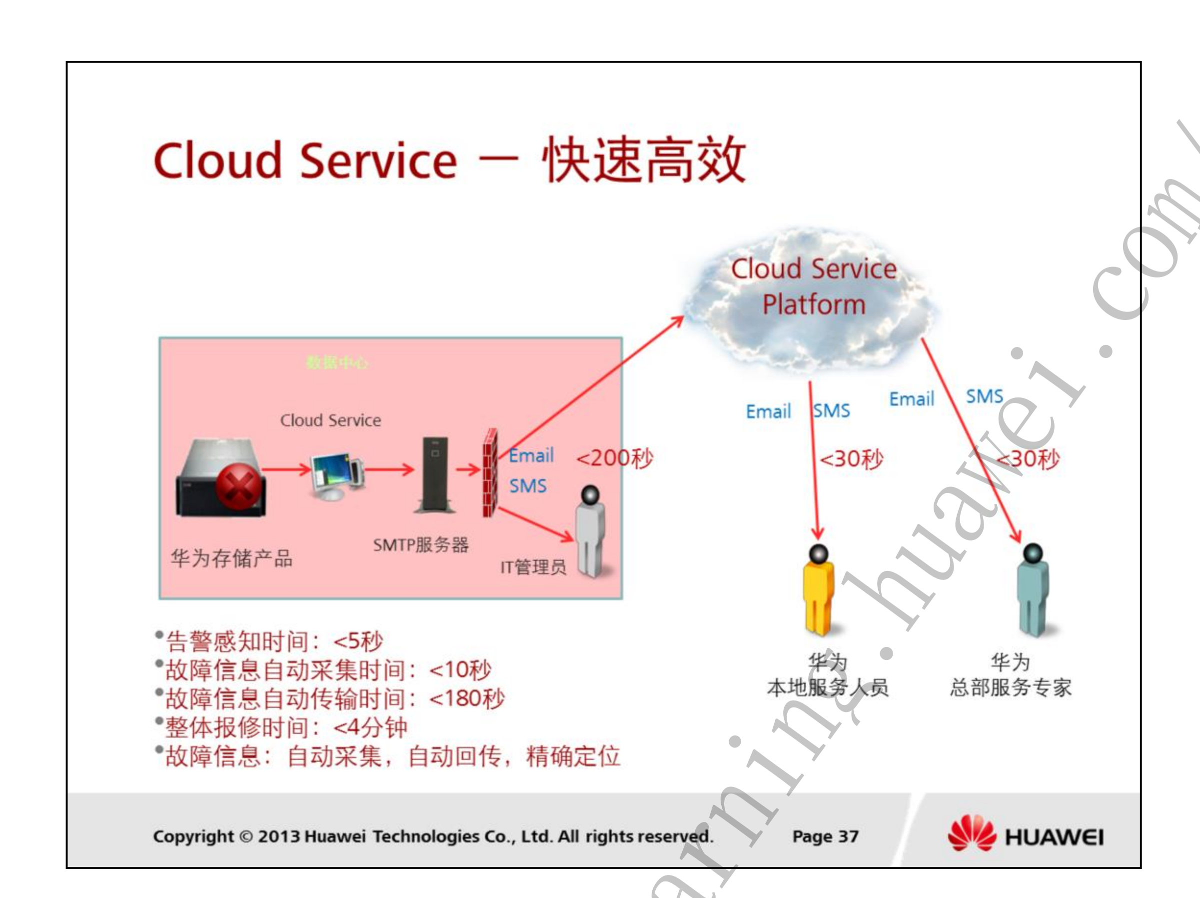
  部署功能可以帮助技术服务工程师、运维工程师进行设备初始化。包括线缆检测工具和硬盘扫描工具。

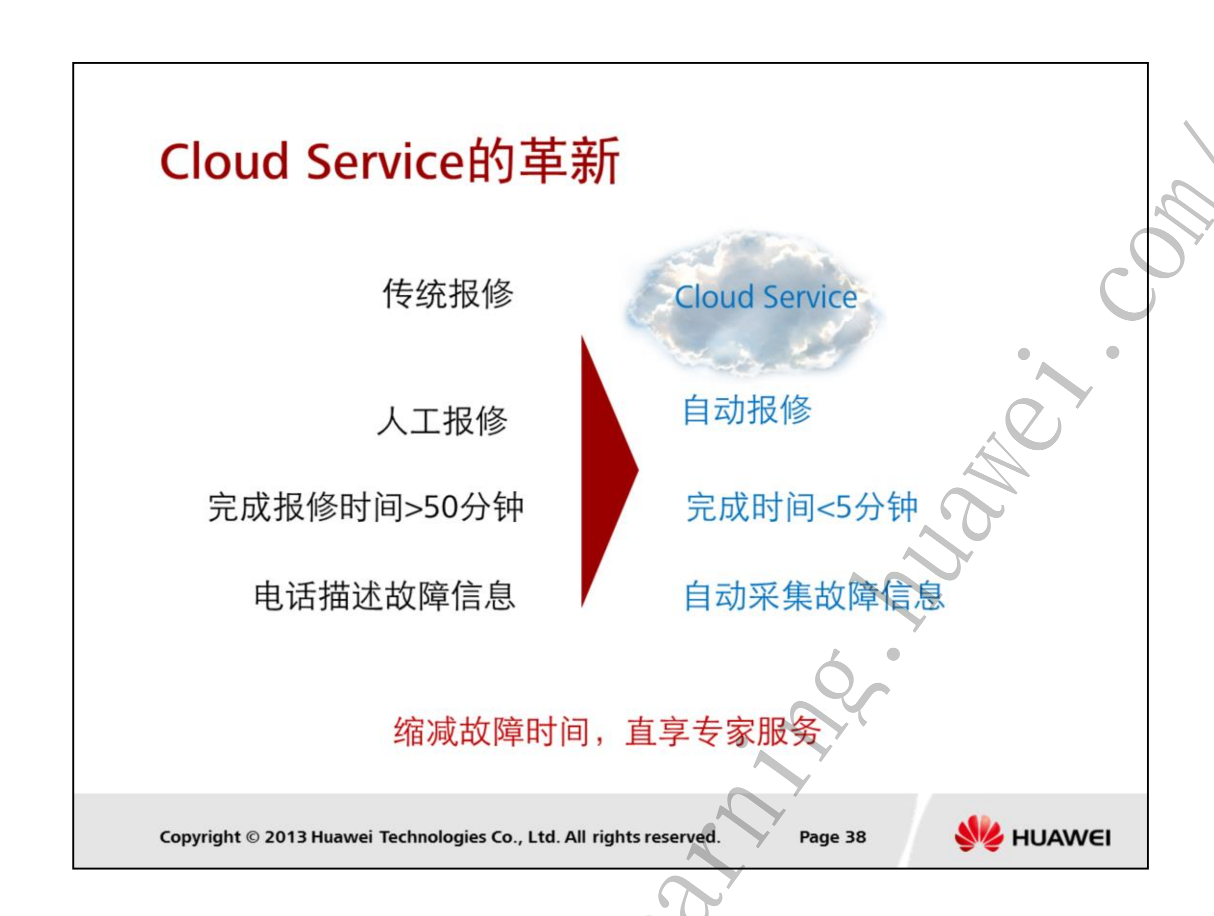
- 部署功能

  部署功能可以帮助技术服务工程师、运维工程师进行设备初始化。包括线缆检测工具和硬盘扫描工具。

- 升级功能

  升级功能可以帮助技术服务工程师、运维工程师对设备进行升级、扩容。包括设备升级工具和扩容工具


Cloud Service — 快速高效
Cloud Service Platform
Cloud Service
Email
SMS
<200秒
华为存储产品
SMTP服务器
IT管理员
Email
SMS
<30秒
Email
SMS
<30秒
华为
本地服务人员
华为
总部服务专家
•告警感知时间：<5秒
•故障信息自动采集时间：<10秒
•故障信息自动传输时间：<180秒
•整体报修时间：<4分钟
•故障信息：自动采集，自动回传，精确定位
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 37
HUAWEI


Cloud Service的革新
传统报修
Cloud Service
人工报修
自动报修
完成报修时间>50分钟
完成时间<5分钟
电话描述故障信息
自动采集故障信息
缩减故障时间，直享专家服务
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 38
HUAWEI

Cloud Service介绍 — 四大特点

主动健康检查

- 存储设备全面定期检查；发生故障时，触发深度健康检查

告警即时感知

- 24*7监控设备告警，告警感知时间小于5秒钟

自动故障通报

- 告警达到规定级别，发送邮件至云服务平台；云服务平台自动转发邮件到相关人员邮箱，并短信告知

故障信息即时回传

- 故障检查报告自动收集，回传到华为总部支持中心，总部服务专家远程故障分析



Cloud Service介绍－客户侧



- 24*7全天候监控
- 定时健康检查
- 潜在故障检查
- 故障触发健康检查

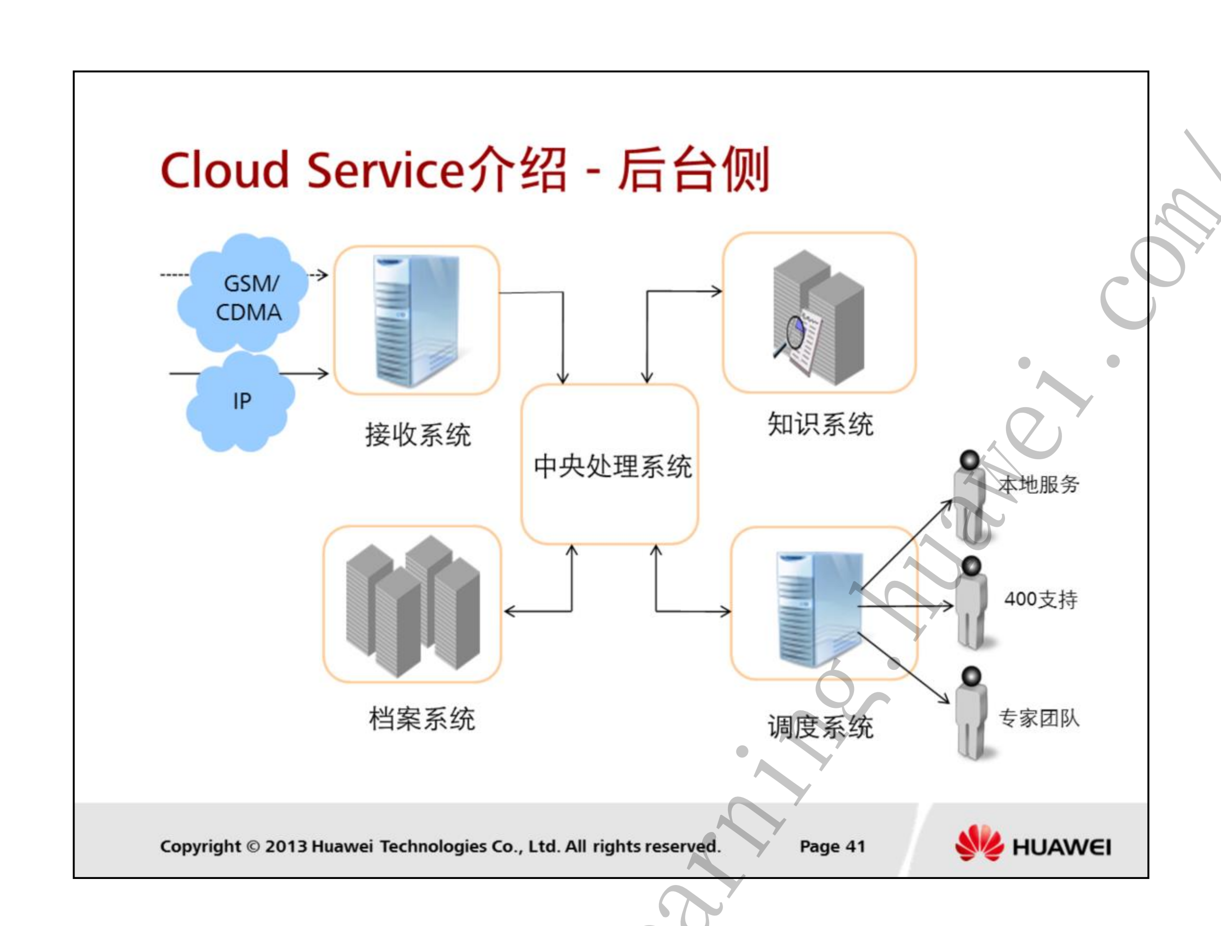


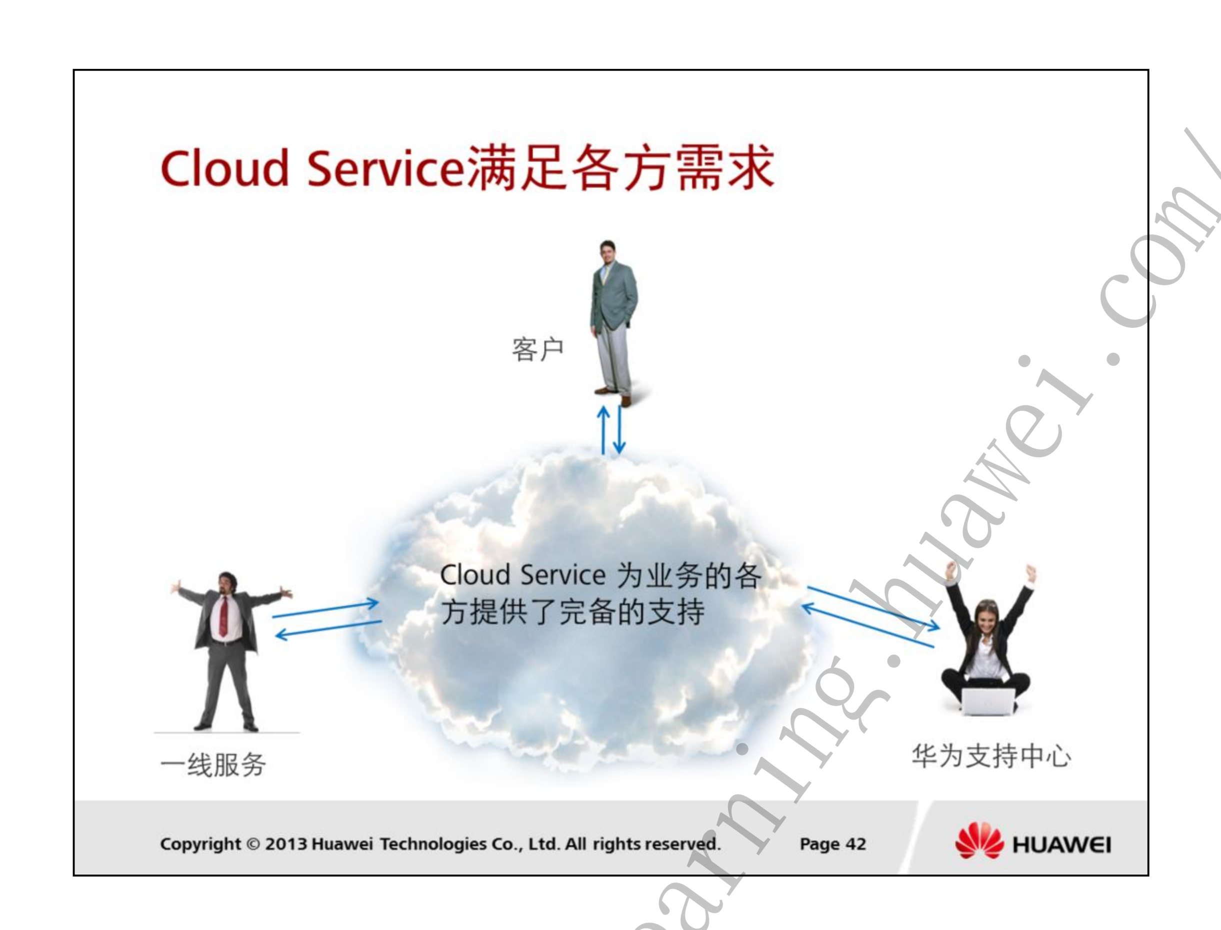
- 采用公有端口
- 仅传送设备运行信息，不传送业务数据
- 是否向外发送，用户完全控制
- 传送信息用户可审计
- 用户信息AES128加密

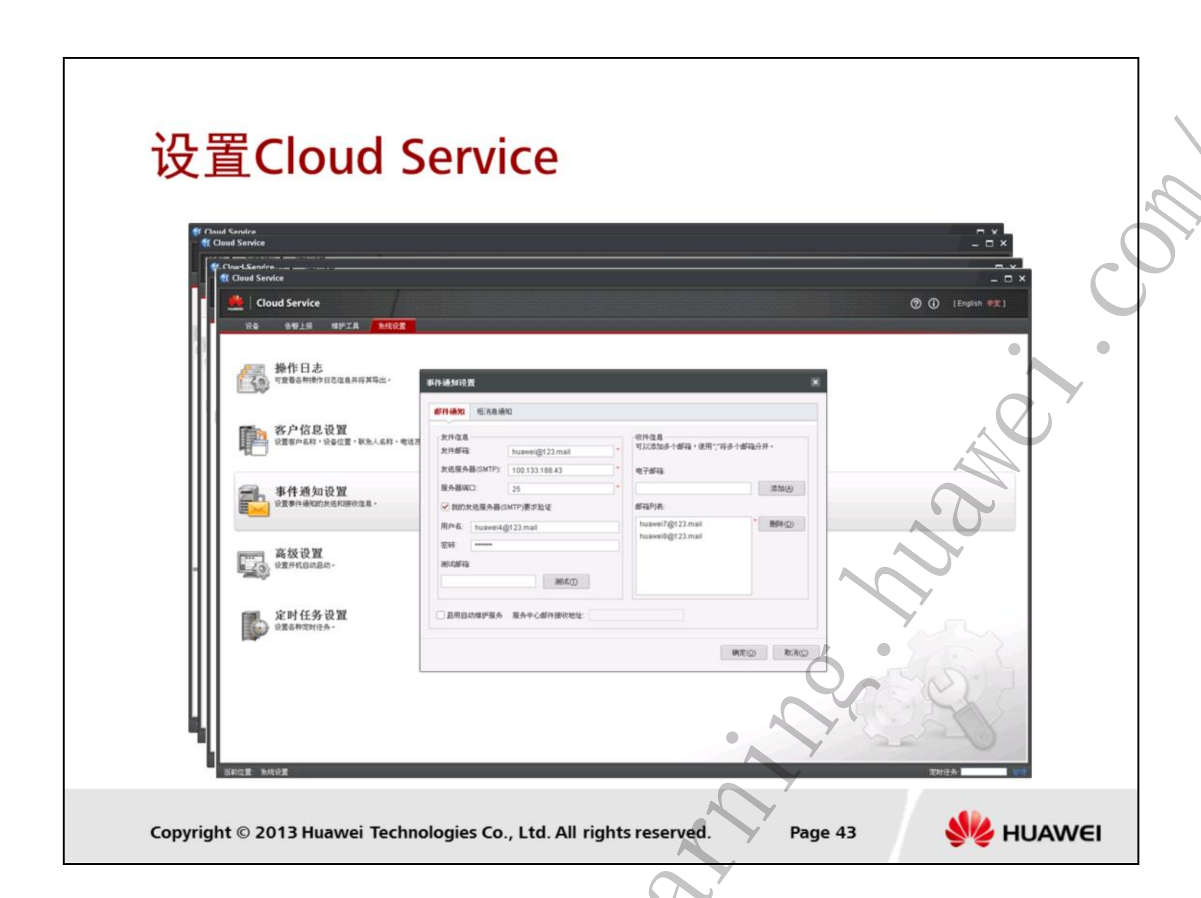


- 支持Windows主力版本
- 绿色软件，解压即用
- 操作简单，所见即所得




Cloud Service介绍 - 后台侧
GSM/
CDMA
IP
接收系统
中央处理系统
知识系统
档案系统
调度系统
本地服务
400支持
专家团队
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 41
HUAWEI


Cloud Service满足各方需求
客户
Cloud Service 为业务的各方提供了完备的支持
一线服务
华为支持中心
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 42
HUAWEI


设置Cloud Service
Cloud Service
操作日志
客户信息设置
事件通知设置
高级设置
定时任务设置
事件通知设置
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 43
HUAWEI



目录

1. 产品硬件结构介绍
2. 硬件安装和连线
3. 勘测和包装拆卸
4. 维护工具
5. **兼容性基础**



存储兼容性相关组件



存储的组网环境非常复杂，主要表现在涉及兼容性的组件及各个组件的型号比较繁多。

后续将分别针对这些组件介绍其影响兼容性结论的相关点。

- 操作系统
- 主机HBA
- 交换机
- 多路径软件
- 上层应用软件（卷管理和集群软件）

 Page 45



操作系统

涉及的主要OS类型

- ✓ Windows：2003/2008 …
- ✓ Linux：SUSE/Red Hat/Red Flag …
- ✓ IBM AIX：AIX 5.3/6.1/7.1 …
- ✓ Oracle Solaris：Solaris 8/9/10/11 …
- ✓ HP-UX：11i v1/v2/v3 …
- ✓ Mac OS：10.5.x/10.6.x/10.7.x …
- ✓ 虚拟机：VMware、XenServer及其它

兼容性结果影响点

操作系统内核

定制化操作系统的内核可能与原版内核会有所不同

操作系统版本

CPU架构

x86,x86_64,SPARC,itanium等不同的架构其兼容性结果不同

 Page 46



主机HBA

涉及的HBA卡类型及参数

- ✓主机接口：iSCSI、FC、SAS
- ✓原产厂商：FC（QLogic、Emulex、LSI、Brocade、ATTO等）；iSCSI（QLogic等）；SAS（LSI、Adaptec、ATTO等）
- ✓OEM厂商：IBM、HP、DELL、SUN、HUAWEI
- ✓插槽类型：PCI、PCI-X、PCI-E
- ✓速率：1Gbps/2Gbps/3Gbps/4Gbps/8Gbps/ 10Gbps 等

兼容性结果影响点



HUAWEI

交换机

涉及的交换机类型及参数

✓主机接口：iSCSI、FC、SAS、10GE

✓原产厂商：

iSCSI（Cisco、HUAWEI）、

FC（Brocade、QLogic、Cisco）、

SAS（IBM、HP、Dell）、

10GE（Huawei、Cisco等）

✓OEM厂商：FC（IBM、HP、DELL、HUAWEI等）

✓速率：1Gbps/2Gbps/4Gbps/8Gbps/10Gbps

兼容性结果影响点

交换机
交换机型号
内部配置参数
Firmware版本
拓扑模式与速率

多路径

涉及的主要多路径类型

- ✓ 自研多路径：Windows /Linux /AIX （iSCSI接口的个别系统及一些非主流操作系统暂无自研多路径支持）
- ✓ 第三方多路径：HP PVLinks、NMP/Solaris STMS/Mac ATTO/Linux DMP/VMware NMP）

兼容性结果影响点

多路径类型

负载均衡模式

多路径类型
不同的多路径在设定路径优先级、路径切换等方面不同

负载均衡模式
不同的负载均衡模式兼容性结论不同

 Page 49 HUAWEI

上层应用软件（卷管理和集群软件）

涉及的主要 集群软件

- ✓ Windows：MSCS、WSFC
- ✓ AIX：HACMP
- ✓ Solaris：Sun Cluster
- ✓ HP-UX：MC/SG
- ✓ Linux：RedHat(RHCS)
- ✓ 第三方：SF、Oracle RAC、ROSE、 HeartBeat

兼容性结果影响点

集群软件版本：操作系统和集群软件版本上不存在一一对应的限制

集群软件类型：不同类型的集群软件差异很大

集群预留模式：主要存在下面几种预留模式：
- •非预留
- •SCSI-2预留
- •SCSI-3预留

 Page 50 HUAWEI

如何获取和使用存储兼容性列表



存储兼容性列表是存储兼容性实验室最重要的输出件。

如何获取并查看该输出件对项目支撑及项目维护来说至关重要！




获取方法一
登陆华为3MS共享平台：
http://3ms.huawei.com/
HOME
DOCUMENTS
CASE
CONNECT
Minisite
My 3MS
Corporate Level
By Product
by Unit
By Region
Competitive Info
Sales Product List (SPL)
Multimedia
I-SHARE
I-VIDEO
Software
Core Network Convergent Communication
IT Solutions
Mobile Network Products
Energy & Infrastructure
Fixed Network products
HuaweiSymantec
Terminal
Service Product & Solution
Hisilicon
Regional MKTG Materials
More
Cloud Data Center
Cloud Platform
Desktop Cloud
Server
Storage
Media Network
Brand
IT General Strategy
Business Model
在上面的网页中，依次选择：DOCUMENTS\By Product\IT Solutions\Storage 进入存储系统资料库。
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 52
HUAWEI


获取方法一
在存储资料库网页左侧的导航栏中，依次选择：
Overall\Bidding\Compatibility List
Directory > Storage > Overall > Bidding > Compatibility List
By Product
Search
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 53
HUAWEI


获取方法二
通过下面的链接直接进入上述目录：
http://3ms.huawei.com/mm/docNav/mmNavigate.do?method=showMMList&node_id=1-2-35621-39226-39244-39256
Copyright © 2013 Huawei Technologies Co., Ltd. All rights reserved.
Page 54
HUAWEI

使用兼容性列表

以S5500T的兼容性列表为例，详细介绍一下兼容性列表的目录结构，如下所示：

FC连接：

FC组网时的组合信息。

iSCSI连接：

iSCSI组网时的组合信息。

带内管理软件：

主机带内管理软件兼容性信息。

HostAgent：

主机客户端兼容性信息。

ISM：

ISM管理软件的兼容性信息。

兼容性认证：

第三方宣称支持华为存储的信息。





使用兼容性列表

通常，当收到一个普通的兼容性需求的时候，可以按照右边的步骤检查兼容性列表。

只有右边条件都满足的时候，才说明该需求是已经经过验证的。否则请咨询相关的接口人。

选择组网
（FC连接还是iSCSI连接）

基本连通性查询
（OS和HBA等是否兼容）

交换机查询
（如果是交换机组网）

卷管理查询
(如果对卷管理有要求)

集群软件查询
（如果存在集群应用）

结束

使用兼容性列表

注意事项

✓在一些项目需求中（多数为投标的项目），可能会对第三方认证信息等有要求，此时可以在兼容性认证目录中进行查询。

✓如果项目中的操作系统为虚拟机，则在虚拟机目录中查询，不在基本连通性中。

✓对于卷管理这块内容，我们只体现第三方的卷管理软件，对于系统自带的卷管理软件，默认测试通过，在兼容性列表中不单独体现。



求助渠道

对于存储兼容性列表以外的内容，可以按照项目的方式，根据**《存储项目兼容性需求反馈模版》**，收集详细项目配置信息，通过产品项目接口人反馈详细的兼容性需求，协助答复。

需求反馈渠道：

售前项目：各产品售前项目支持接口同事

售后项目：各产品交付维护组接口同事



Thank you

www.huawei.com

华为职业认证通过者权益

通过**任一项**华为职业认证，您即可在华为在线学习网站(*http://learning.huawei.com/cn*) 享有如下特权：

- 1、**华为E-learning 课程学习**
  - **内容**：**所有华为职业认证E-Learning课程**，扩展您在其他技术领域的技术知识
  - **方式**：请提交您的“华为账号”和注册账号的“email地址”到 *Learning@huawei.com 申请权限。*
- 2、**华为培训教材下载**
  - **内容：华为职业认证培训教材+华为产品技术培训教材，**覆盖企业网络、存储、安全等诸多领域
  - **方式**：登录*华为在线学习网站*，进入“*华为培训->面授培训*”，在具体课程页面即可下载教材。
- 3、**华为在线公开课(LVC)优先参与**
  - **内容**：企业网络、UC&C、安全、存储等诸多领域的职业认证课程，华为讲师授课，开班人数有限
  - **方式**：开班计划及参与方式请详见LVC排期：http://support.huawei.com/learning/NavigationAction!createNavi#navi[id]=_16
- 4、**学习工具 eNSP**
  - *eNSP (Enterprise Network Simulation Platform)*，是由华为提供的免费的、可扩展的、图形化网络仿真工具。主要对企业网路由器和交换机进行硬件模拟，完美呈现真实设备实景；同时也支持大型网络模拟，让大家在没有真实设备的情况下也能够进行实验测试。
- 另外，华为建立了知识分享平台 华为认证论坛。**您可以在线与华为技术专家交流技术，与其他考生分享考试经验，一起学习华为产品技术。**（http://support.huawei.com/ecommunity/bbs/list_2247.html）

